Percezione come Inferenza Variazionale

Fondamenti Epistemici, Struttura Matematica e Implicazioni per la Filosofia della Mente

Radici Epistemologiche della Percezione come Inferenza

La questione di come il sistema nervoso trasformi il caotico flusso di segnali fisici in un'esperienza coerente del mondo non è semplicemente una domanda empirica risolvibile accumulando dati neurofisiologici. Essa tocca uno dei nodi più profondi dell'intera tradizione filosofica occidentale: il rapporto tra la struttura del soggetto conoscente e la struttura di ciò che viene conosciuto. Che la percezione non sia un'acquisizione passiva di informazioni bensì un processo attivo di costruzione e interpretazione è un'intuizione che precede di gran lunga la scienza cognitiva contemporanea, affondando le radici nell'empirismo critico di Kant e nella fisiologia speculativa del tardo Ottocento. Il merito della ricerca neuroscientifica degli ultimi decenni non consiste tanto nell'aver scoperto questa verità quanto nell'averla resa suscettibile di trattazione quantitativa rigorosa, traducendola in un formalismo matematico il cui potere esplicativo supera di molto quello delle formulazioni originarie.

Hermann von Helmholtz, nella sua monumentale opera di ottica fisiologica pubblicata tra il 1856 e il 1867, coniò l'espressione *unbewusster Schluss* — inferenza inconscia — per descrivere la natura della percezione visiva. L'argomento di Helmholtz era epistemologicamente sofisticato: i dati sensoriali che raggiungono la retina sono radicalmente sottodeterminati rispetto alle cause fisiche che li producono. Una distribuzione di luminanza sulla retina è compatibile con un'infinità di configurazioni tridimensionali di oggetti, sorgenti luminose e proprietà di riflessione. Il sistema visivo non dispone dunque di accesso diretto alle cause del segnale; deve invece operare una forma di ragionamento inferenziale, seppure del tutto automatico e inaccessibile all'introspezione cosciente, che risale dalle conseguenze osservate alle cause probabilmente responsabili. Helmholtz era abbastanza avveduto da riconoscere la struttura logica di tale processo: si tratta di un'abductio nel senso peirciano, di una inferenza alla spiegazione migliore, mediata da conoscenze pregresse incorporate nel sistema nervoso attraverso l'esperienza passata e, plausibilmente, attraverso l'adattamento filogenetico.

La parentela con l'epistemologia kantiana è immediata e non meramente metaforica. Kant aveva sostenuto nella *Critica della ragion pura* che l'intelletto non si limita a ricevere passivamente il materiale dell'intuizione sensibile, ma lo struttura attivamente attraverso le categorie dell'intelletto puro — causalità, sostanza, necessità — senza le quali nessuna esperienza sarebbe possibile. Queste forme a priori della conoscenza non sono proiettate sull'esperienza in modo arbitrario; esse costituiscono piuttosto le condizioni di possibilità di qualsiasi esperienza coerente. In Helmholtz questo schema trascendentale trova una naturalizzazione parziale: le "aspettative" incorporate nel sistema visivo svolgono una funzione analoga alle categorie kantiane, strutturando il segnale sensoriale secondo schemi di regolarità appresi, ma il loro statuto non è più quello di condizioni pure e immutabili bensì di distribuzioni di probabilità modificabili dall'esperienza. Il kantismo diviene, in questa lettura, un kantismo empirico e probabilistico — che è precisamente ciò che la visione computazionale contemporanea ha reso tecnico.

Il passaggio dalla formulazione helmholtziana a quella rigorosamente bayesiana richiede di apprezzare la profondità del problema dell'inversione che la percezione deve risolvere. Si parla comunemente di "problema dell'ottica inversa" per indicare la difficoltà computazionale fondamentale: la formazione delle immagini è un processo diretto che proietta uno spazio tridimensionale ricco su una superficie bidimensionale povera, distruggendo informazione in modo irreversibile. Il processo percettivo deve invertire questa proiezione, ricostruendo le cause a partire dagli effetti. Ora, qualsiasi inversione di questo tipo è matematicamente mal posta nel senso di Hadamard: non esiste soluzione unica, la soluzione non dipende in modo continuo dai dati, e piccole perturbazioni nei segnali sensoriali possono corrispondere a enormi variazioni nelle scene interpretate. La percezione stabile e veridica non è dunque un fatto banale che si spiega da sé; essa richiede che il sistema nervoso adotti vincoli aggiuntivi, ovvero incorpori conoscenze a priori sulla distribuzione delle cause plausibili nel mondo.

Questa è precisamente la struttura del ragionamento bayesiano. Se chiamiamo h un'ipotesi percettiva — una descrizione della scena tridimensionale — e d i dati sensoriali disponibili, il teorema di Bayes prescrive:

$$P(h \mid d) = P(d \mid h) \cdot P(h) / P(d)$$

La distribuzione posteriore $P(h|d)$ rappresenta il grado di credenza razionale nell'ipotesi h alla luce dei dati d. Essa è il prodotto della verosimiglianza $P(d|h)$ — quanto sono probabili questi dati se l'ipotesi h fosse vera, ovvero il modello generativo diretto — per la distribuzione a priori $P(h)$ — la conoscenza preesistente sulla plausibilità delle varie ipotesi — normalizzata dalla verosimiglianza marginale $P(d)$, che funge da costante di normalizzazione. La percezione ottimale, in questa prospettiva, è un processo di inferenza bayesiana che computa o approssima la distribuzione posteriore $P(h|d)$ e ne estrae, ad esempio, il massimo a posteriori (MAP) come stima percettiva puntuale, oppure usa l'intera distribuzione per guidare le decisioni e le azioni.

Il programma di ricerca che va sotto il nome di "cervello bayesiano" — sviluppato sistematicamente da Weiss, Simoncelli e Adelson, da Kersten, Mamassian e Yuille, da Knill e Pouget nei primi anni Duemila — ha mostrato con impressionante dettaglio quantitativo che molti fenomeni percettivi sistematicamente bizzarri o controintuitivi trovano spiegazione elegante come inferenza bayesiana con priori fisiologicamente plausibili. Le illusioni visive diventano, in questa lettura, non errori del sistema ma inferenze ottimali sotto priori errati o in situazioni artificiali che violano le regolarità statistiche dell'ambiente naturale. L'illusione di movimento dei gradini di Adelson-Bergen, la percezione di profondità dalle ombre, l'integrazione multimodale audiovisiva: tutti questi fenomeni mostrano una firma quantitativa della predizione bayesiana, con varianze combinate che seguono la regola della somma pesata degli inversi delle varianze — la regola di combinazione ottimale degli stimatori indipendenti sotto perdita quadratica.

Tuttavia, affermare che il cervello "fa inferenza bayesiana" senza ulteriori specificazioni è quasi vuoto di contenuto computazionale. Il problema serio non è normativo — sappiamo che l'inferenza bayesiana è razionalmente ottimale — bensì algoritmico e implementativo. Come potrebbe un sistema neurale calcolare effettivamente la distribuzione posteriore su spazi di stati ad altissima dimensionalità? La risposta diretta è: non può farlo, almeno non in generale. Il calcolo esatto della distribuzione a posteriori richiede la valutazione dell'integrale

$$P(d) = \int P(d \mid h) \cdot P(h)\, dh$$

che per modelli generativi non banali è analiticamente intrattabile e computazionalmente proibitivo. È proprio questa considerazione che rende indispensabile il passaggio all'inferenza variazionale — il cuore matematico del formalismo che si intende analizzare nelle sezioni seguenti.

Prima di compiere questo passaggio, conviene sostare un momento sulla riflessione di Richard Gregory, che negli anni Settanta e Ottanta sviluppò in modo indipendente e con linguaggio meno tecnico ma filosoficamente ricco un'analisi della percezione come formulazione e verifica di ipotesi. Per Gregory, le percezioni sono letteralmente ipotesi sul mondo esterno — congetture azzardate sulla base di dati incompleti, esattamente come le teorie scientifiche sono congetture sulla struttura della realtà. Le illusioni percettive sono ipotesi sbagliate che il sistema adotta perché corrispondono alla spiegazione statisticamente più probabile in condizioni normali. Questa tesi, enunciata da Gregory con eleganza letteraria ma senza apparato formale, acquisisce tutta la sua potenza esplicativa solo quando viene inscritta nel framework dell'inferenza variazionale e del principio di energia libera, dove i termini "ipotesi", "probabilità" e "spiegazione migliore" ricevono una definizione matematica precisa.

Vi è infine una considerazione di ordine filosofico più generale che merita di essere introdotta prima di addentrarsi nel formalismo. La prospettiva della percezione come inferenza bayesiana ha implicazioni radicali per il problema classico dello scetticismo epistemologico. Se il sistema percettivo opera essenzialmente come un motore di inferenza che costruisce rappresentazioni del mondo a partire da segnali ambigui

mediante l'applicazione di conoscenze pregresse, allora l'accesso diretto al mondo esterno — il "realismo ingenuo" che fonda il senso comune — è una finzione funzionale piuttosto che una verità metafisica. Ciò che sperimentiamo fenomenologicamente come presenza immediata del mondo è, su questa concezione, la superficie cosciente di un processo computazionale profondo e opaco che ha già eseguito l'inferenza e ne presenta il risultato — l'ipotesi a posteriori più probabile — come se fosse il mondo stesso. L'esperienza percettiva è dunque, nel senso tecnico del termine, un'allucinazione controllata: una proiezione dall'interno che corrisponde, nella grande maggioranza dei casi ecologicamente rilevanti, alle strutture causali del mondo esterno, ma che può divergere da esse in condizioni artificiali o patologiche.

Questa prospettiva non è esente da difficoltà filosofiche. La più immediata riguarda la circolarità apparente: se le aspettative a priori sono formate dall'esperienza passata, e l'esperienza passata è essa stessa il prodotto di inferenze percettive precedenti, come si garantisce che il processo nel suo insieme converga verso rappresentazioni veridiche del mondo piuttosto che verso sistemi chiusi di aspettative auto-confermanti? La risposta che emerge dal framework formale è che il sistema non può garantire la veridicità assoluta delle sue rappresentazioni, ma può minimizzare l'errore di predizione — la discrepanza tra i segnali sensoriali previsti e quelli effettivamente ricevuti — e che questa minimizzazione, in un ambiente naturale sufficientemente stazionario, tende a produrre modelli interni che rispecchiano le regolarità strutturali del mondo. La corrispondenza tra modello interno e mondo esterno è dunque un risultato statistico atteso di un processo adattativo, non una garanzia metafisica a priori. Il realismo, in questa prospettiva, non è un fondamento ma una conquista — sempre provvisoria, sempre suscettibile di revisione.

II. Fondamenti Matematici dell'Inferenza Variazionale

II.1 — L'intrattabilità della posterioree la necessità di un'approssimazione

La struttura formale del problema percettivo, come si è visto, richiede il calcolo della distribuzione a posteriori $P(h|d)$ su un insieme di ipotesi h dati i segnali sensoriali d. Prima di affrontare le tecniche di approssimazione, è opportuno rendere esplicita la fonte dell'intrattabilità computazionale, poiché essa non è accidentale ma strutturale e rivela qualcosa di profondo sulla natura dei sistemi percettivi biologici.

Si consideri un modello generativo gerarchico in cui lo stato del mondo $\psi \in \mathbb{R}^n$ è un vettore ad alta dimensionalità che genera segnali sensoriali $s \in \mathbb{R}^m$ attraverso una funzione di verosimiglianza $P(s|\psi)$. La distribuzione marginale dei segnali sensoriali è:

$$P(s) = \int_{\mathbb{R}^n} P(s \mid \psi)\, P(\psi)\, d\psi$$

Questo integrale, detto *verosimiglianza marginale* o *evidenza del modello*, è analitico solo per classi molto ristrette di modelli — essenzialmente le coppie gaussiane-lineari, dove sia il modello generativo che la distribuzione a priori sono gaussiane e la funzione di verosimiglianza è lineare. Fuori da questo caso, che rappresenta un'idealizzazione radicale dell'effettiva complessità del mondo, l'integrale non ha forma chiusa. Nei modelli gerarchici profondi, dove ogni livello trasforma nonlinearmente l'attivazione del livello sottostante — come accade nei modelli che aspirano a descrivere la corteccia visiva reale — l'intrattabilità è la norma, non l'eccezione.

L'inferenza variazionale risponde a questa difficoltà sostituendo il problema di calcolo integrabile con un problema di ottimizzazione. L'idea centrale è di introdurre una famiglia parametrica di distribuzioni $Q(\psi; \mu)$, indicizzata da parametri $\mu \in \mathcal{V}$ appartenenti a uno spazio variazionale opportunamente scelto, e di trovare il membro di questa famiglia che meglio approssima la vera distribuzione a posteriori $P(\psi \mid s)$. Il criterio di "migliore approssimazione" è la divergenza di Kullback-Leibler.

II.2 — La divergenza di Kullback-Leibler e le sue proprietà fondamentali

La divergenza KL tra due distribuzioni di probabilità Q e P su uno spazio Ω è definita come:

$$D_{KL}(Q \,\|\, P) = \int_{\Omega} Q(\psi) \ln \left[Q(\psi) \,/\, P(\psi) \right] d\psi$$

Per distribuzioni discrete su un insieme finito, l'integrale diviene somma:

$$D_{KL}(Q \,\|\, P) = \sum_i Q(\psi_i) \ln \left[Q(\psi_i) \,/\, P(\psi_i) \right]$$

Le proprietà della divergenza KL meritano una trattazione accurata perché sono direttamente rilevanti per l'interpretazione del formalismo variazionale. La proprietà più importante è la *non-negatività*, che discende direttamente dalla disuguaglianza di Jensen applicata alla funzione strettamente convessa $f(t) = -\ln(t)$:

$$\begin{aligned} D_{KL}(Q \,\|\, P) &= E_Q[\, \ln(Q/P) \,] = -E_Q[\, \ln(P/Q) \,] \\ &\geq -\ln E_Q[\, P/Q \,] \qquad \text{(Jensen: } -\ln \text{ è convessa)} \\ &= -\ln \int Q \cdot (P/Q)\, d\psi \\ &= -\ln \int P\, d\psi \\ &= -\ln 1 = 0 \end{aligned}$$

con uguaglianza se e solo se $Q(\psi) = P(\psi)$ quasi ovunque. La divergenza KL è dunque una misura di dissimilarità tra distribuzioni, sempre non-negativa e nulla sse le distribuzioni coincidono. Non è però una metrica in senso stretto: non è simmetrica

— in generale $D_{KL}(Q\|P) \neq D_{KL}(P\|Q)$ — né soddisfa la disuguaglianza triangolare.

L'asimmetria della KL ha conseguenze pratiche profonde che il sistema percettivo implicitamente deve "scegliere" di gestire. La divergenza $KL(Q\|P)$ — detta anche KL *forward* o *inclusiva* — penalizza fortemente i casi in cui Q assegna massa di probabilità positiva laddove P è vicino a zero; tende pertanto a produrre distribuzioni approssimanti Q che si "allargano" per coprire tutti i modi della distribuzione bersaglio P. La divergenza $KL(P\|Q)$ — KL *reverse* o *esclusiva* — penalizza invece i casi in cui Q è vicino a zero laddove P ha massa significativa, tendendo a produrre approssimazioni Q concentrate attorno al modo dominante di P. Il framework dell'inferenza variazionale standard, come si vedrà, minimizza $D_{KL}(Q\|P)$, con le conseguenze epistemologiche che ciò comporta.

II.3 — L'Evidence Lower BOund (ELBO) e l'energia libera variazionale

Il punto di partenza algebrico è la decomposizione esatta della log-verosimiglianza marginale $\ln P(s)$. Si moltiplica e si divide l'integrando per $Q(\psi; \mu)$:

$$
\begin{aligned}
\ln P(s) &= \ln \int P(s, \psi)\, d\psi \\
&= \ln \int Q(\psi; \mu) \cdot [\, P(s, \psi) \,/\, Q(\psi; \mu) \,]\, d\psi \\
&\geq \int Q(\psi; \mu) \cdot \ln [\, P(s, \psi) \,/\, Q(\psi; \mu) \,]\, d\psi \qquad \text{(Jensen, poiché ln è concava)} \\
&= E_{Q}[\, \ln P(s, \psi) \,] - E_{Q}[\, \ln Q(\psi; \mu) \,] \\
&\equiv \mathcal{L}(\mu; s)
\end{aligned}
$$

La quantità $\mathcal{L}(\mu; s)$ è l'Evidence Lower BOund, o ELBO. Essa costituisce un limite inferiore della log-evidenza del modello, e l'uguaglianza è raggiunta soltanto quando $Q(\psi; \mu) = P(\psi \mid s)$ esattamente. Il gap tra $\ln P(s)$ e $\mathcal{L}$ è una quantità non negativa che si può calcolare esplicitamente:

$$
\begin{aligned}
&\ln P(s) - \mathcal{L}(\mu; s) \\
&= \ln P(s) - [\, E_Q[\ln P(s,\psi)] - E_Q[\ln Q(\psi;\mu)] \,] \\
&= E_Q[\ln P(s)] - E_Q[\ln P(s,\psi)] + E_Q[\ln Q(\psi;\mu)] \\
&= E_Q[\ln P(s) - \ln P(s,\psi) + \ln Q(\psi;\mu)] \\
&= E_Q[\ln Q(\psi;\mu) - \ln P(\psi \mid s)] \\
&= D_{KL}(\, Q(\psi;\mu) \,\|\, P(\psi \mid s) \,)
\end{aligned}
$$

Si ottiene dunque la decomposizione esatta e fondamentale:

$$\ln P(s) = \mathcal{L}(\mu; s) + D_{KL}(\, Q(\psi;\mu) \,\|\, P(\psi \mid s) \,)$$

Poiché il lato sinistro è fisso rispetto a μ, massimizzare l'ELBO $\mathcal{L}$ rispetto a μ equivale a minimizzare $D_{KL}(Q\|P(\psi \mid s))$, cioè ad avvicinare la distribuzione variazionale Q alla vera distribuzione a posteriori. Questo è il risultato centrale dell'inferenza variazionale:

il problema intractable di calcolare $P(\psi | s)$ è sostituito da un problema di ottimizzazione su μ, che può essere affrontato con algoritmi di gradiente.

L'energia libera variazionale F è definita come il negativo dell'ELBO:

$$F(\mu; s) = -\mathcal{L}(\mu; s) = E_Q[\ \ln Q(\psi; \mu) - \ln P(s, \psi)\ ]$$

Si può riscrivere questa espressione in due modi illuminanti sotto profili diversi. Il primo è la decomposizione in termini di verosimiglianza attesa e divergenza a priori:

$$F(\mu; s) = -E_Q[\ \ln P(s | \psi)\ ] + D_KL(\ Q(\psi;\mu) \parallel P(\psi)\ )$$

Il primo termine è la *verosimiglianza negativa attesa* — misura quanto bene le predizioni del modello si accordano con i dati sensoriali s, in media rispetto alla distribuzione variazionale Q sui possibili stati nascosti ψ. Il secondo termine è la divergenza KL tra la distribuzione variazionale e la distribuzione a priori — misura quanto Q si discosta dalle aspettative pregresse. Questa decomposizione svela immediatamente la struttura di compromesso (trade-off) implicita nell'inferenza variazionale: minimizzare F significa trovare un equilibrio tra la fedeltà ai dati (ridurre l'errore di predizione) e la parsimonia rispetto al modello pregresso (non allontanarsi troppo dal a priori). Questa tensione è l'analogo formale della coppia kantiana intuizione-concetto, o del principio di parsimonia epistemologica noto come rasoio di Occam.

Il secondo modo di riscrivere F utilizza la nozione di entropia di Shannon. L'entropia differenziale di Q è:

$$H[Q] = -E_Q[\ \ln Q(\psi; \mu)\ ] = -\int Q(\psi; \mu) \ln Q(\psi; \mu)\, d\psi$$

Allora:

$$\begin{aligned} F(\mu; s) &= E_Q[\ -\ln P(s, \psi)\ ] - H[Q] \\ &= E_Q[\ -\ln P(s | \psi) - \ln P(\psi)\ ] + H[Q]^{-1} \cdot (-H[Q]) \end{aligned}$$

oppure, più chiaramente:

$$F(\mu; s) = E_Q[\ -\ln P(s, \psi)\ ] - H[Q]$$

che esprime F come differenza tra l'energia media del sistema (nel senso statistico-meccanico, dove l'energia è l'autoformazione negativa del log-congiunto) e l'entropia della distribuzione variazionale. Questa formulazione rende trasparente il legame con la meccanica statistica: la definizione di energia libera di Helmholtz in termodinamica è $F = \langle E \rangle - TS$, dove $\langle E \rangle$ è l'energia media, T la temperatura e S l'entropia. Il formalismo variazionale riproduce esattamente questa struttura con $T = 1$ (temperature unitaria), il che non è una coincidenza formale ma riflette connessioni

profonde tra la teoria dell'informazione, la meccanica statistica e la teoria della computazione.

II.4 — Famiglie variazionali e l'approssimazione di campo medio

Il problema variazionale non è completamente specificato finché non si sceglie la famiglia parametrica $Q(\psi; \mu)$ entro cui effettuare l'ottimizzazione. La scelta di questa famiglia determina la qualità dell'approssimazione e la trattabilità del problema di ottimizzazione risultante. Esistono due grandi classi di approssimazioni variazionali, entrambe rilevanti per la modellazione neurale.

L'approssimazione di *campo medio* — o approssimazione di indipendenza fattorizzata — assume che la distribuzione variazionale si fattorizzi sui diversi componenti del vettore di stati nascosti:

$$Q(\psi; \mu) = \Pi_{i=1}^{n} Q_i(\psi_i; \mu_i)$$

ovvero che le diverse dimensioni dello spazio degli stati siano trattate come statisticamente indipendenti sotto la distribuzione approssimante. Questa assunzione è quasi certamente falsa per la vera distribuzione a posteriori, che generalmente presenta correlazioni tra componenti diverse. Tuttavia, produce un considerevole vantaggio computazionale: il problema di ottimizzazione si decompone in una sequenza di problemi disaccoppiati, ciascuno dei quali aggiorna una singola Q_i tenendo fisse le altre. L'aggiornamento ottimale della j-esima componente, derivato minimizzando F rispetto a Q_j, ha la forma:

$$\ln Q_j^*(\psi_j) = E_{Q_{i\neq j}}[\ln P(s, \psi)] + \text{cost}$$

dove $E_{Q_{i\neq j}}$ indica il valore atteso rispetto a tutte le distribuzioni variazionali eccetto la j-esima, e "cost" è la costante di normalizzazione. Questo produce un sistema di equazioni di punto fisso che si risolve iterativamente — ogni fattore Q_j viene aggiornato in sequenza usando le aspettative correnti dei fattori complementari — procedimento noto come inferenza variazionale di campo medio o algoritmo di aggiornamento coordinato.

La connessione con la neurobiologia è suggestiva: l'algoritmo di campo medio assomiglia strutturalmente ai meccanismi di passaggio di messaggi (message passing) sulle reti bayesiane, e diversi autori hanno proposto che i circuiti corticali ricorrenti implementino varianti di questo schema, dove le attivazioni neuronali codificano i parametri delle distribuzioni Q_i e le sinapsi ricorrenti effettuano gli scambi di aspettative tra le diverse componenti.

II.5 — L'approssimazione di Laplace

La seconda grande famiglia di approssimazioni variazionali, particolarmente rilevante per l'analisi della codifica predittiva gerarchica, è l'approssimazione di Laplace. Essa

assume che la distribuzione variazionale sia gaussiana, centrata sul massimo a posteriori $\hat{\psi} = \text{argmax}_\psi \ln P(s, \psi)$, con matrice di precisione pari all'opposto dell'hessiano del log-congiunto nel punto di massimo:

$$Q(\psi; \mu) = \mathcal{N}(\psi; \hat{\psi}, \Sigma)$$

dove:

$$\hat{\psi} = \text{argmax}_\psi \left[\ln P(s|\psi) + \ln P(\psi) \right]$$

$$\Sigma = -\left[\partial^2/\partial\psi^2 \ln P(s, \psi) \,|_{\{\psi=\hat{\psi}\}} \right]^{-1}$$

L'approssimazione di Laplace è giustificata dall'espansione di Taylor del log-congiunto attorno al suo massimo. Definendo $\Phi(\psi) \equiv \ln P(s, \psi)$, si ha:

$$\Phi(\psi) \approx \Phi(\hat{\psi}) + (\psi - \hat{\psi})^T \nabla\Phi|_{\{\hat{\psi}\}} + (1/2)(\psi - \hat{\psi})^T H|_{\{\hat{\psi}\}} (\psi - \hat{\psi})$$

dove $\nabla\Phi|\{\hat{\psi}\} = 0$ *per definizione di massimo, e* $H|\{\hat{\psi}\} = \partial^2\Phi/\partial\psi^2|_{\{\hat{\psi}\}}$ è l'hessiano nel punto di massimo, che è una matrice negativa semidefinita (poiché $\hat{\psi}$ è un massimo). Dunque:

$$P(s, \psi) \approx P(s, \hat{\psi}) \cdot \exp\left[(1/2)(\psi - \hat{\psi})^T H|_{\{\hat{\psi}\}} (\psi - \hat{\psi}) \right]$$

che è proporzionale a una distribuzione gaussiana con media $\hat{\psi}$ e matrice di covarianza $\Sigma = (-H|_{\{\hat{\psi}\}})^{-1}$. L'energia libera variazionale sotto l'approssimazione di Laplace diviene:

$$\begin{aligned} F_{\text{Laplace}} &= -\ln P(s, \hat{\psi}) - (1/2) \ln \det(2\pi\Sigma) \\ &= -\ln P(s, \hat{\psi}) - (n/2) \ln(2\pi) + (1/2) \ln \det(-H|_{\{\hat{\psi}\}}) \end{aligned}$$

Il contributo entropico alla riduzione dell'energia libera è interamente catturato dal termine $(1/2) \ln \det(2\pi\Sigma)$, che cresce al diminuire della curvatura della distribuzione a posteriori — ovvero al crescere dell'incertezza sull'ipotesi percettiva più probabile. Questo è un aspetto cruciale: sotto l'approssimazione di Laplace, il sistema non cerca semplicemente di massimizzare la verosimiglianza a posteriori nel senso puntuale (MAP), ma bilancia questa con la "larghezza" della distribuzione, premiando le configurazioni in cui l'ipotesi è allo stesso tempo molto verosimile e ragionevolmente precisa.

II.6 — Il gradiente dell'ELBO e la dinamica dell'aggiornamento percettivo

Il passo finale nel collegare il formalismo variazionale alla dinamica neurale è derivare le equazioni di aggiornamento per i parametri variazionali μ. Sotto l'approssimazione di Laplace con modelli generativi gerarchici gaussiani, questo porta direttamente alle equazioni della codifica predittiva. Ma prima di specializzare al caso gaussiano, è istruttivo considerare la forma generale del gradiente dell'ELBO rispetto ai parametri variazionali.

Si scriva la distribuzione variazionale come $Q(\psi; \mu)$ e si assuma che essa appartenga a una famiglia esponenziale con statistiche sufficienti $T(\psi)$ e parametri naturali $\eta(\mu)$:

$$Q(\psi; \eta) = h(\psi) \exp[\, \eta(\mu)^T T(\psi) - A(\eta) \,]$$

dove $A(\eta)$ è il logaritmo della funzione di partizione, che garantisce la normalizzazione. Il gradiente dell'ELBO rispetto ai parametri naturali η è:

$$\partial \mathcal{L} / \partial \eta = E_Q[\, T(\psi) \cdot (\ln P(s,\psi) - \ln Q(\psi;\eta)) \,] + \mathrm{Cov}_Q[T(\psi), \ln P(s,\psi) - \ln Q(\psi;\eta)]$$

Questa espressione, pur nella sua generalità, rivela una struttura importante: il gradiente dell'ELBO è legato alla covarianza tra le statistiche sufficienti $T(\psi)$ e l'errore di predizione logaritmico $\ln P(s,\psi) - \ln Q(\psi;\eta)$. In termini neuroscientifici: l'aggiornamento delle rappresentazioni interne (i parametri μ) è guidato dalla correlazione tra l'attivazione corrente dei neuroni che codificano $T(\psi)$ e l'errore di predizione. Questa è precisamente la struttura della regola di apprendimento di Hebb che emerge naturalmente dall'ottimizzazione variazionale — un risultato di notevole interesse per la neurobiologia teorica.

Nel caso specifico dei modelli gaussiani lineari gerarchici — che costituiscono il modello di riferimento per la codifica predittiva nel senso di Rao e Ballard — la struttura si semplifica considerevolmente. Si consideri un modello a due livelli con:

$$s = g(v) + \varepsilon_s, \quad \varepsilon_s \sim \mathcal{N}(0, \Sigma_s)$$
$$v = f(u) + \varepsilon_v, \quad \varepsilon_v \sim \mathcal{N}(0, \Sigma_v)$$

dove $s \in \mathbb{R}^m$ sono le attivazioni sensoriali, $v \in \mathbb{R}^k$ è uno stato nascosto al primo livello, $u \in \mathbb{R}^l$ è uno stato nascosto al secondo livello, e $g: \mathbb{R}^k \to \mathbb{R}^m$, $f: \mathbb{R}^l \to \mathbb{R}^k$ sono funzioni di mappatura (potenzialmente nonlineari). Le matrici Σ_s e Σ_v sono le covarianze (inverse delle precisioni) del rumore ai due livelli.

Sotto l'approssimazione di Laplace, gli stati nascosti vengono stimati minimizzando l'energia libera, che per questo modello diviene:

$$F = (1/2)[\, \varepsilon^T \Pi_s \varepsilon + \varepsilon_v^T \Pi_v \varepsilon_v \,] + \text{cost}$$

dove $\Pi_s = \Sigma_s^{-1}$ e $\Pi_v = \Sigma_v^{-1}$ sono le matrici di precisione (inverse delle covarianze), e gli errori di predizione sono:

$\varepsilon = s - g(\hat{v})$ (errore sensoriale)
$\varepsilon_v = \hat{v} - f(\hat{u})$ (errore di livello superiore)

Il gradiente dell'energia libera rispetto a $\hat{v}$ — che governa la dinamica dell'aggiornamento della stima del primo livello — è:

$$\partial F/\partial \hat{v} = -(\partial g/\partial \hat{v})^T \Pi_s \varepsilon + \Pi_v \varepsilon_v$$

Questa equazione ha una lettura neurobiologica immediata: l'aggiornamento della rappresentazione $\hat{v}$ è guidato da due segnali contrastanti. Il primo termine, $-(\partial g/\partial \hat{v})^T \Pi_s \varepsilon$, è un segnale ascendente (bottom-up) che porta l'errore di predizione sensoriale pesato dalla precisione; il secondo termine, $\Pi_v \varepsilon_v$, è un segnale discendente (top-down) che porta l'errore di predizione rispetto al livello superiore, anch'esso pesato dalla precisione. Il sistema converge quando questi due segnali si bilanciano — cioè quando la stima $\hat{v}$ è al contempo compatibile con i dati sensoriali e con le aspettative generate dal livello gerarchico superiore.

L'equazione di aggiornamento in tempo continuo, ottenuta come gradiente di discesa sull'energia libera, è:

$$d\hat{v}/dt = -(\partial g/\partial \hat{v})^T \Pi_s \varepsilon + \Pi_v \varepsilon_v$$

$$= \Pi_v [\hat{v} - f(\hat{u})] - (\partial g/\partial \hat{v})^T \Pi_s [s - g(\hat{v})]$$

Questa equazione differenziale ordinaria descrive la dinamica di un sistema che minimizza continuamente la propria energia libera, elaborando simultaneamente segnali ascendenti e discendenti pesati per la rispettiva incertezza. È notevole che questa struttura emergga necessariamente dalla minimizzazione dell'energia libera variazionale senza alcuna assunzione ad hoc sulla connettività neurale: la gerarchia discendente-ascendente è una conseguenza della struttura del modello generativo gerarchico, non un presupposto.

La ponderazione per la precisione (Π_s ed Π_v) gioca un ruolo cruciale che va oltre la semplice combinazione ottimale di segnali. Le precisioni possono esse stesse essere parametri della distribuzione variazionale, soggetti a ottimizzazione. Un sistema che apprende e aggiorna le proprie stime di precisione è in grado di regolare dinamicamente il peso relativo attribuito ai segnali ascendenti e discendenti — ovvero di modulare la propria "apertura" all'informazione sensoriale rispetto alla propria "fiducia" nelle predizioni interne. Si tratta di un meccanismo formalmente preciso che, come si argomentà nella parte successiva dedicata al principio di energia libera, potrebbe rendere conto di fenomeni come l'attenzione, la previsione temporale e alcune forme di psicopatologia percettiva.

III. Geometria dell'Informazione e Struttura Riemanniana dello Spazio delle Distribuzioni

---

III.1 — Lo spazio delle distribuzioni come varietà differenziale

Il formalismo variazionale sviluppato nella sezione precedente lascia in ombra una questione che ha implicazioni sia pratiche che concettuali di primo piano: quale struttura geometrica è naturalmente associata allo spazio delle distribuzioni di probabilità entro cui si compie l'ottimizzazione? La risposta a questa domanda, elaborata sistematicamente da Shun-ichi Amari a partire dagli anni Sessanta e raccolta nella teoria nota come *geometria dell'informazione*, rivela che lo spazio delle distribuzioni è una varietà riemanniana con una metrica canonicamente determinata — non per convenzione ma per ragioni di invarianza profonde — e che le traiettorie di ottimizzazione variazionale acquisiscono un significato geometrico preciso in questo spazio. Ignorare questa struttura e trattare il problema variazionale come ottimizzazione in uno spazio euclideo piatto produce algoritmi subottimali e interpretazioni fuorvianti.

Sia $\mathscr{S} = \{ P(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^n \}$ una famiglia parametrica di distribuzioni di probabilità su uno spazio campionario $\mathscr{X}$, con parametri $\theta = (\theta^1, \theta^2, \ldots, \theta^n)$. Si assume che la mappa $\theta \mapsto P(\cdot;\theta)$ sia iniettiva e sufficientemente regolare da permettere le operazioni di differenziazione richieste. In queste condizioni, $\mathscr{S}$ è una varietà differenziale di dimensione n, e i parametri θ costituiscono un sistema di coordinate locali. Il punto cruciale è che esistono molti sistemi di coordinate equivalenti per descrivere la stessa famiglia — per esempio, una reparametrizzazione $\theta \mapsto \eta(\theta)$ biunivoca e differenziabile produce un sistema altrettanto valido. Una nozione geometrica genuina deve essere invariante rispetto a queste reparametrizzazioni.

Sia $\ell(x;\theta) = \ln P(x;\theta)$ il log-likelihood. Il vettore score al punto θ è il vettore gradiente:

$$\partial \ell(x;\theta) = \partial \ln P(x;\theta) / \partial\theta^l \qquad \text{per } l = 1, \ldots, n$$

Il valore atteso dello score sotto $P(\cdot;\theta)$ è identicamente nullo — è questo il contenuto della cosiddetta prima identità di Bartlett:

$$E_{P(\cdot;\theta)}[\,\partial \ell(x;\theta)\,] = \int [\partial\, P(x;\theta)]\, dx = \partial \int P(x;\theta)\, dx = \partial\, 1 = 0$$

(dove lo scambio tra derivata e integrale è giustificato da condizioni di regolarità standard). Segue che la matrice di covarianza dello score è ben definita:

$$g\ (\theta) = E_{P(\cdot;\theta)}[\,\partial \ell \cdot \partial \ell\,] = \mathrm{Cov}_{P(\cdot;\theta)}[\,\partial \ell, \partial \ell\,]$$

Questa matrice di covarianza è la *matrice di informazione di Fisher*, e il tensore di rango (0,2) con componenti g (θ) definisce una metrica riemanniana su $\mathscr{S}$, la *metrica di Fisher*. Si può mostrare che g ammette una rappresentazione alternativa tramite la seconda identità di Bartlett:

g (θ) = -E_{P(·;θ)}[∂ ∂ ℓ(x; θ)]

cioè la matrice di Fisher è l'opposto dell'hessiano medio del log-likelihood — la curvatura media della funzione log-verosimiglianza, con segno cambiato. Questa formulazione la collega direttamente all'approssimazione di Laplace discussa nella sezione II.5: la matrice di covarianza $\Sigma = (-H|_{\hat{\psi}})^{-1}$ dell'approssimazione gaussiana è proprio l'inversa della matrice di Fisher valutata nel massimo a posteriori.

III.2 — Invarianza della metrica di Fisher e unicità

La proprietà che eleva la metrica di Fisher da semplice scelta tra infinite metriche possibili a struttura geometrica canonica è la sua *invarianza per trasformazioni statisticamente sufficienti*. Una statistica $T: \mathscr{X} \to \mathscr{Y}$ è sufficiente per θ se la distribuzione di T(x) contiene tutta l'informazione rilevante su θ contenuta in x — ovvero se P(x | T(x), θ) è indipendente da θ. Il teorema di Chentsov (1972) afferma che la metrica di Fisher è, a meno di un fattore scalare positivo, l'unica metrica riemanniana su $\mathscr{S}$ che sia monotona (ovvero non-aumentante) sotto mappature markoviane stocastiche. In termini operativi: elaborare i dati in modo stocastico non può aumentare l'informazione di Fisher — un risultato che è l'analogo geometrico del data-processing inequality della teoria dell'informazione.

La conseguenza per il problema variazionale è immediata e profonda. Quando si discende il gradiente dell'energia libera nello spazio euclideo dei parametri μ — usando l'aggiornamento ordinario:

μ_new = μ_old - α · ∇_μ F(μ)

si sta implicitamente misurando le distanze nello spazio dei parametri con la metrica euclidea, che è arbitraria e dipende dalla parametrizzazione scelta. Un passaggio da θ a $\eta = \eta(\theta)$ cambierà la forma delle curve di discesa e, in generale, rallenterà o accelererà la convergenza in modi che non hanno significato statistico. La discesa di gradiente naturale di Amari corregge questa arbitrarietà usando la metrica di Fisher come metrica dello spazio:

$$\mu_{new} = \mu_{old} - \alpha \cdot G(\mu)^{-1} \nabla_\mu F(\mu)$$

dove G(μ) è la matrice di Fisher valutata in μ e $G(\mu)^{-1} \nabla_\mu F$ è il *gradiente naturale* — il gradiente dell'energia libera espresso nella metrica riemanniana intrinseca dello spazio delle distribuzioni. Il gradiente naturale è invariante per reparametrizzazione: se si passa da μ a $\nu = \nu(\mu)$, il gradiente naturale si trasforma in modo covariante e le

traiettorie di discesa sono geometricamente le stesse, indipendentemente dalla parametrizzazione.

In pratica, la discesa di gradiente naturale converge molto più rapidamente di quella euclidea per i problemi di inferenza variazionale tipici, perché tiene conto della curvatura dello spazio delle distribuzioni. Nei modelli con struttura di ridondanza parametrica — situazione comune nelle reti neurali profonde — il gradiente euclideo tende a compiere passi inefficienti in direzioni dove la curvatura è bassa, mentre il gradiente naturale adatta automaticamente la dimensione del passo alla struttura locale della varietà.

III.3 — Connessioni duali e geometria α

La geometria dell'informazione rivela una struttura ancora più ricca: lo spazio $\mathscr{S}$ è dotato non di una ma di un'intera famiglia di connessioni affini, parametrizzate da un numero reale $\alpha \in \mathbb{R}$, denominate *connessioni a* o connessioni di Amari-Chentsov. Per ogni α, si definisce la connessione α-affine tramite i suoi simboli di Christoffel:

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = E_\theta\left[\left(\partial_i \partial_j \ell + \frac{1-\alpha}{2} \cdot \partial_i \ell \cdot \partial_j \ell\right) \cdot \partial_k \ell \right]$$

I due casi più importanti sono $\alpha = 1$ (connessione esponenziale, o e-connessione) e $\alpha = -1$ (connessione mista, o m-connessione). La coppia di connessioni $(\nabla^{(1)}, \nabla^{(-1)})$ è *duale* rispetto alla metrica di Fisher, nel senso che soddisfa la relazione:

$$\partial_k\, g(\partial_i, \partial_j) = g(\nabla^{(1)}_{\partial_k} \partial_i, \partial_j) + g(\partial_i, \nabla^{(-1)}_{\partial_k} \partial_j)$$

Le geodetiche rispetto alla e-connessione sono le *rette esponenziali* — curve della forma $P(x;\theta(t)) = P(x;\theta_0) \cdot \exp[t \cdot c(x)] \,/\, Z(t)$, dove $Z(t)$ è la costante di normalizzazione. Le geodetiche rispetto alla m-connessione sono le *rette delle miscele* — curve della forma $P(x;\theta(t)) = (1-t)\, P(x;\theta_0) + t\, P(x;\theta_1)$.

La rilevanza di questa struttura per l'inferenza variazionale è che la divergenza KL può essere espressa come una sorta di "quadrato della distanza" in termini delle geodetiche duali. In particolare, il teorema di Pitagora generalizzato di Amari afferma che se M è una sottovarietà e-affine (chiusa rispetto a combinazioni esponenziali) e N è una sottovarietà m-affine di $\mathscr{S}$, e se P^* è il punto di intersezione di una e-geodetica da $P \in N$ a M con una m-geodetica da $Q \in M$ a N, allora:

$$D_{KL}(P \,\|\, Q) = D_{KL}(P \,\|\, P^*) + D_{KL}(P^* \,\|\, Q)$$

Questo è l'esatto analogo del teorema di Pitagora euclideo per le distanze: la "distanza informazionale" (KL) tra P e Q si decompone nella somma delle distanze da P alla proiezione su M e dalla proiezione a Q, quando le due proiezioni sono effettuate rispetto alle connessioni duali. Questo teorema formalizza la struttura dell'inferenza

variazionale: trovare la distribuzione in Q più vicina (nel senso KL) alla vera a posteriori è una proiezione m-affine della vera a posteriori sulla varietà variazionale Q.

III.4 — La metrica di Fisher nei modelli gaussiani e il ruolo della precisione

Per i modelli gaussiani — rilevanti per la codifica predittiva — la metrica di Fisher assume una forma esplicita che merita di essere calcolata esplicitamente. Si consideri la famiglia delle distribuzioni gaussiane univariate $\mathcal{N}(\mu, \sigma^2)$, parametrizzate da $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. Il log-likelihood è:

$$\ell(x; \mu, \sigma^2) = -(1/2)\ln(2\pi\sigma^2) - (x-\mu)^2/(2\sigma^2)$$

Le derivate prime sono:

$$\partial\ell/\partial\mu = (x - \mu)/\sigma^2$$
$$\partial\ell/\partial\sigma^2 = -(1/(2\sigma^2)) + (x-\mu)^2/(2\sigma^4)$$

La matrice di Fisher è dunque:

$$g_{11} = E[\,(\partial\ell/\partial\mu)^2\,] = E[\,(x-\mu)^2/\sigma^4\,] = \sigma^2/\sigma^4 = 1/\sigma^2$$

$$\begin{aligned} g_{22} &= E[\,(\partial\ell/\partial\sigma^2)^2\,] \\ &= E[\,(-(1/(2\sigma^2)) + (x-\mu)^2/(2\sigma^4))^2\,] \\ &= (1/4)E[\,(1/\sigma^2 - (x-\mu)^2/\sigma^4)^2\,] \\ &= (1/4)\cdot[\,1/\sigma^4 - 2/\sigma^6 \cdot E[(x-\mu)^2] + 1/\sigma^8 \cdot E[(x-\mu)^4]\,] \\ &= (1/4)\cdot[\,1/\sigma^4 - 2/\sigma^4 + 3/\sigma^4\,] \\ &= (1/4)\cdot 2/\sigma^4 = 1/(2\sigma^4) \end{aligned}$$

$$\begin{aligned} g_{12} &= E[\,(\partial\ell/\partial\mu)(\partial\ell/\partial\sigma^2)\,] \\ &= E[\,(x-\mu)/\sigma^2 \cdot (-(1/(2\sigma^2)) + (x-\mu)^2/(2\sigma^4))\,] \\ &= (1/(2\sigma^6))\,E[(x-\mu)^3] - (1/(2\sigma^4))\,E[(x-\mu)] \\ &= 0 - 0 = 0 \end{aligned}$$

(dove si è usato il fatto che i momenti dispari di una gaussiana centrata sono nulli). La metrica di Fisher sulla varietà delle gaussiane univariate è pertanto diagonale:

$$G(\mu, \sigma^2) = \mathrm{diag}(1/\sigma^2, 1/(2\sigma^4))$$

Il termine $g_{11} = 1/\sigma^2 = \Pi$ (la precisione) è particolarmente significativo: misura la curvatura della varietà nella direzione della media, ed è inversamente proporzionale alla varianza. In altri termini, la metrica di Fisher "vede" le distribuzioni con varianza piccola come più distanti tra loro (per una data differenza di media) rispetto alle distribuzioni con varianza grande. Questo è intuitivamente corretto: distinguere due

gaussiane strette con medie vicine è statisticamente più difficile che distinguere due gaussiane larghissime con le stesse medie — la prima situazione è "più critica" statisticamente, e la metrica di Fisher la riflette assegnando maggiore distanza geometrica.

Per i modelli gaussiani multivariati $\mathcal{N}(\mu, \Sigma)$ con $\mu \in \mathbb{R}^n$ e Σ matrice di covarianza $n \times n$ definita positiva, la metrica di Fisher nella direzione dei parametri di media è data dal blocco:

$$G_{\mu\mu} = \Sigma^{-1} = \Pi$$

la matrice di precisione. Nella direzione dei parametri di covarianza, la metrica ha una forma più complessa che coinvolge il prodotto di Kronecker delle matrici di precisione — il cosiddetto *prodotto di Khatri-Rao* — ma il punto cruciale rimane: la geometria dello spazio delle distribuzioni gaussiane è completamente governata dalla struttura di precisione del modello, e l'aggiornamento variazionale ottimale in senso geometrico è quello che usa la precisione come metrica.

Questo risultato riversa luce nuova sul ruolo della precisione nelle equazioni della codifica predittiva derivate nella sezione II.6. La ponderazione degli errori di predizione per le matrici di precisione Π e Π_v non è soltanto una scelta di combinazione ottimale degli stimatori: essa è l'espressione del gradiente naturale dell'energia libera sullo spazio delle distribuzioni gaussiane. Il sistema di codifica predittiva, cioè, esegue implicitamente una discesa di gradiente naturale — non euclideo — nella sua dinamica di aggiornamento percettivo. Questo potrebbe spiegare perché questi sistemi convergono in modo robusto e rapidamente in molti contesti, nonostante la nonlinearità dei modelli generativi reali.

III.5 — Distanza di Wasserstein e geometria ottimale del trasporto

Accanto alla geometria dell'informazione fondato sulla KL e sulla metrica di Fisher, negli ultimi quindici anni si è sviluppata una teoria alternativa ma complementare basata sulla geometria del *trasporto ottimale*, che assegna allo spazio delle distribuzioni una struttura riemanniana formalmente distinta ma in certi contesti più naturale. La distanza di Wasserstein di ordine 2 tra due distribuzioni P e Q su $\mathbb{R}^n$ è definita come:

$$W_2(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \left[\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \, d\gamma(x, y) \right]^{1/2}$$

dove $\Gamma(P, Q)$ è l'insieme dei *piani di trasporto* — distribuzioni congiunte su $\mathbb{R}^n \times \mathbb{R}^n$ con marginali P e Q rispettivamente. La distanza W_2 misura il "costo minimo" di trasportare la massa di probabilità da P a Q, dove il costo di spostare una unità di massa dal punto x al punto y è $\|x-y\|^2$.

La geometria indotta da W_2 è profondamente diversa da quella indotta dalla metrica di Fisher. In particolare, la metrica di Wasserstein è sensibile alla struttura spaziale del supporto delle distribuzioni — due distribuzioni molto simili come funzioni ma con

supporti geometricamente distanti sono lontane nella metrica W_2, mentre possono essere vicine nella metrica di Fisher. Questa differenza non è accidentale: le due geometrie catturano aspetti complementari della struttura probabilistica, e la loro integrazione è un problema aperto di notevole profondità.

Per l'inferenza variazionale, la geometria di Wasserstein ha guadagnato interesse come base per la definizione di *flussi di particelle* — algoritmi in cui un insieme di particelle che approssima la distribuzione corrente Q viene evoluto seguendo la discesa di gradiente nella metrica W_2, anziché aggiornando parametri di una distribuzione parametrica. In questo framework, la discesa di gradiente di Wasserstein dell'energia libera KL(Q‖P) produce l'equazione differenziale alle derivate parziali:

$$\partial Q/\partial t = \nabla \cdot [\, Q \cdot \nabla (\ln Q - \ln P) \,]$$

che è l'equazione di Fokker-Planck associata alla dinamica di Langevin:

$$d\psi = -\nabla_\psi [\, \ln Q(\psi) - \ln P(\psi) \,]\, dt + \sqrt{2}\, dW$$

dove dW è un processo di Wiener (moto browniano) in $\mathbb{R}^n$. Questa connessione tra la discesa di gradiente variazionale e la dinamica stocastica di Langevin è un risultato di Jordan, Kinderlehrer e Otto (1998) di grande portata: rivela che i processi diffusivi stocastici nei sistemi fisici e biologici possono essere interpretati come processi di minimizzazione dell'energia libera nella geometria di Wasserstein. Nel contesto neurale, ciò suggerisce che il rumore neuronale — tradizionalmente trattato come fastidio da eliminare — potrebbe svolgere un ruolo funzionale nella dinamica di campionamento e inferenza approssimata, un'idea sviluppata nella teoria della *codifica predittiva campionante* di Fiser, Berkes e Lengyel.

IV. Il Principio di Energia Libera: dall'Inferenza all'Azione, dagli Stati Markoviani all'Esistenza

---

IV.1 — La generalizzazione fristoniana: oltre la percezione passiva

Il formalismo dell'inferenza variazionale, nella sua formulazione originaria, descrive un sistema che riceve segnali sensoriali e aggiorna passivamente le proprie rappresentazioni interne per minimizzare l'energia libera. Karl Friston, a partire dal 2005 e con crescente elaborazione nei decenni successivi, ha proposto una generalizzazione radicale di questo schema che aspira a costituire un principio unificante per la biologia, la cognizione e l'azione: il *principio di energia libera*. Nella sua versione più ambiziosa, questo principio afferma che qualsiasi sistema fisico che mantenga la propria integrità rispetto a un ambiente perturbante deve, necessariamente e per ragioni matematicamente costrittive, minimizzare una quantità

formalmente equivalente all'energia libera variazionale. La percezione e l'azione diventano allora due modalità complementari attraverso cui un sistema biologico persegue questa minimizzazione.

Il punto di partenza della generalizzazione è una considerazione di carattere probabilistico-termodinamico. Un organismo vivente occupa, nello spazio degli stati fisiologici e comportamentali, una regione relativamente ristretta — quella compatibile con la vita. Questa regione può essere caratterizzata da una distribuzione di probabilità stazionaria $P^*(s)$ sui segnali sensoriali che l'organismo "si aspetta" di ricevere in condizioni normali. La sopravvivenza equivale al mantenimento della distribuzione sensoriale effettiva $Q(s)$ vicina a questa distribuzione stazionaria attesa $P^*(s)$. In termini informativi, la sopravvivenza corrisponde a mantenere bassa la *sorpresa* — definita come il log-evidenza negativa:

$$-\ln P^*(s)$$

L'organismo non può calcolare direttamente $-\ln P^*(s)$ perché non conosce la distribuzione marginale $P^*(s)$ in forma chiusa. Tuttavia — e questo è il nucleo del passaggio fristoniano — può sempre calcolare un limite superiore di questa sorpresa, ovvero l'energia libera variazionale:

$$F = D_{KL}(Q(\psi; \mu) \,\|\, P^*(\psi \mid s)) - \ln P^*(s)$$
$$\geq -\ln P^*(s)$$

In quanto $F \geq -\ln P^*(s)$, minimizzare F è una strategia suffuciente per contenere la sorpresa — e dunque per garantire la sopravvivenza. Il sistema non deve quindi fare qualcosa di diverso dall'inferenza variazionale per restare in vita: l'inferenza bayesiana approssimata e la sopravvivenza biologica sono, in questo framework, due descrizioni dello stesso processo.

IV.2 — Il Markov Blanket e la struttura di separazione stocastica

Prima di sviluppare le equazioni dell'azione, è necessario introdurre il concetto formale che consente di distinguere un sistema dal suo ambiente in modo matematicamente preciso: il *Markov blanket*. Il termine, introdotto da Judea Pearl nella teoria dei grafi bayesiani, indica l'insieme dei nodi in una rete bayesiana che "scherma" statisticamente un nodo interno dall'insieme dei nodi esterni: un nodo è condizionalmente indipendente da tutti i nodi esterni dato il suo Markov blanket.

Formalmente, sia $X = (\eta, b, \mu, s)$ un sistema di variabili casuali, dove η sono gli stati esterni (del mondo), $b = (a, \alpha)$ sono gli stati del Markov blanket suddivisi in stati attivi a (che influenzano l'ambiente) e stati sensoriali α (che sono influenzati dall'ambiente), e μ sono gli stati interni del sistema. La struttura di Markov blanket richiede le seguenti indipendenze condizionali:

$$\eta \perp \mu \mid b$$ (gli stati interni sono schermate dagli esterni dato il blanket)

$\mu \perp \eta \mid b$ (simmetrico, per la definizione di blanket)

In un sistema dinamico stocastico descritto da un'equazione di Langevin generalizzata:

$dX/dt = f(X) + \omega$

dove $f: \mathbb{R}^n \to \mathbb{R}^n$ è il flusso deterministico e ω è il rumore, la struttura di Markov blanket è codificata nella struttura di accoppiamento della funzione f. In particolare, la condizione che μ sia schermato da η richiede che $\partial f_\mu / \partial \eta = 0$ — gli stati interni non dipendono direttamente dagli stati esterni, ma solo attraverso gli stati sensoriali del blanket. Analogamente, $\partial f_\eta / \partial \mu = 0$ — gli stati esterni non sono influenzati direttamente dagli stati interni, ma solo attraverso gli stati attivi del blanket.

Questo schema codifica precisamente la struttura di un organismo che interagisce con il proprio ambiente: gli organi sensoriali (stati sensoriali α) trasmettono informazioni dall'ambiente verso l'interno, mentre i muscoli e gli effettori (stati attivi a) trasmettono le conseguenze degli stati interni verso l'ambiente. Gli stati interni μ — corrispondenti al sistema nervoso centrale — non hanno contatto diretto né con le cause profonde del segnale sensoriale né con il mondo fisico, ma solo con il confine del Markov blanket.

IV.3 — L'energia libera su stati generalizzati

Il framework del principio di energia libera diventa particolarmente ricco quando si considerano non soltanto gli stati attuali ma le loro traiettorie temporali. Friston introduce la nozione di *stati generalizzati* — vettori che comprendono lo stato corrente e le sue derivate temporali fino a un ordine arbitrario:

$\tilde{s} = (s, s', s'', ..., s^{(n)})$

dove $s' = ds/dt$, $s'' = d^2s/dt^2$, e così via. L'idea è che il sistema nervoso non elabora soltanto istantanee sensoriali ma integra l'informazione sulla struttura temporale del segnale — la sua velocità, accelerazione, jerk — in un'unica struttura di rappresentazione che cattura il *modo* in cui il segnale si evolve, non soltanto il suo valore corrente.

Il modello generativo su stati generalizzati ha la forma:

$\tilde{s} = \tilde{g}(\tilde{\psi}) + \tilde{\varepsilon}_s$
$\tilde{\psi}' = \tilde{f}(\tilde{\psi}) + \tilde{\varepsilon}_\psi$

dove $\tilde{\psi}$ è il vettore generalizzato degli stati nascosti, $\tilde{g}$ è la funzione di osservazione generalizzata, $\tilde{f}$ è la funzione di transizione generalizzata che governa la dinamica degli stati nascosti, e i termini di rumore $\tilde{\varepsilon}$ sono assunti gaussiani con covarianza $\tilde{\Sigma}$.

La funzione $\tilde{g}$ e $\tilde{f}$ sugli stati generalizzati hanno una struttura triangolare che riflette la struttura delle derivate. Se $g: \mathbb{R}^k \to \mathbb{R}^m$ è la funzione di osservazione al primo ordine, allora:

$$\tilde{g}(\tilde{\psi}) = (g(\psi), g'(\psi, \psi'), g''(\psi, \psi', \psi''), \dots)$$

dove $g'(\psi, \psi') = (\partial g/\partial \psi)\, \psi'$ è la derivata materiale di g lungo la traiettoria $\psi(t)$, e così via per gli ordini superiori. L'energia libera su stati generalizzati è:

$$\tilde{F}(\tilde{\mu}; \tilde{s}) = (1/2)\, \tilde{\varepsilon}\square^{T}\, \tilde{\Pi}\square\, \tilde{\varepsilon}\square + (1/2)\, \tilde{\varepsilon}_\psi^{T}\, \tilde{\Pi}_\psi\, \tilde{\varepsilon}_\psi + (1/2) \ln \det(2\pi\, \tilde{S})$$

dove $\tilde{\varepsilon}\square = \tilde{s} - \tilde{g}(\tilde{\mu})$ e $\tilde{\varepsilon}_\psi = D\, \tilde{\mu} - \tilde{f}(\tilde{\mu})$, con D l'operatore di derivazione temporale sugli stati generalizzati, che agisce spostando ogni componente di un ordine: $D(\psi, \psi', \psi'', \dots) = (\psi', \psi'', \psi''', \dots)$.

Le equazioni di aggiornamento degli stati interni generalizzati $\tilde{\mu}$ si ottengono dal gradiente discendente di $\tilde{F}$:

$$d\tilde{\mu}/dt = D\, \tilde{\mu} - \partial \tilde{F}/\partial \tilde{\mu}$$

Il termine $D\tilde{\mu}$ è cruciale e rappresenta il contributo della struttura temporale: esso "propaga" la stima corrente lungo il suo ordine superiore previsto, prima che l'errore di predizione la corregga. In assenza di errori di predizione, il sistema si limiterebbe a propagare le sue rappresentazioni seguendo la propria dinamica interna predetta — è la correzione $\partial \tilde{F}/\partial \tilde{\mu}$ che forza l'aggiornamento ogni volta che le predizioni sensoriali si discostano dai segnali effettivi.

IV.4 — L'azione come minimizzazione dell'energia libera

Fino a questo punto, il sistema è stato trattato come un osservatore passivo che aggiorna le proprie rappresentazioni interne. L'estensione cruciale del principio di energia libera consiste nell'includere l'azione come ulteriore variabile di ottimizzazione. Gli stati attivi a — le uscite motorie del sistema — compaiono nell'equazione sensoriale perché l'organismo può modificare il proprio ambiente fisico e, così facendo, modificare i segnali sensoriali che riceve. In termini formali, gli stati sensoriali α dipendono sia dagli stati esterni η che dagli stati attivi a:

$$\alpha = \alpha(\eta, a)$$

Il gradiente dell'energia libera rispetto agli stati attivi è:

$$\partial F/\partial a = \partial F/\partial s \cdot \partial s/\partial a = -\varepsilon^{T}\, \Pi\, \cdot \partial g/\partial a$$

L'aggiornamento degli stati attivi per minimizzare l'energia libera è dunque:

$$da/dt = -\partial F/\partial a = \varepsilon^{T} \Pi \cdot \partial s/\partial a$$

Questo ha un'interpretazione elegante e non immediata: le azioni sono guidate dagli errori di predizione sensoriale ε, pesati per la precisione Π e per il jacobiano $\partial s/\partial a$ che descrive come le azioni si traducono in variazioni sensoriali. In altri termini, il sistema agisce in modo da *ridurre la discrepanza tra i segnali sensoriali previsti e quelli ricevuti* — non cercando di modellare il mondo più accuratamente (come nel caso della percezione) ma cambiando il mondo in modo da renderlo più conforme alle proprie previsioni.

Questa formulazione contiene un risultato sorprendente: la distinzione classica tra percezione e azione non è una distinzione tra processi di natura diversa, ma tra due strategie di minimizzazione della stessa funzione. La percezione aggiorna μ per ridurre F tenendo fisso l'ambiente; l'azione aggiorna a per ridurre F tenendo fisse le rappresentazioni interne. La coppia (percezione, azione) è dunque una coppia di strategie coordinate che cooperano nella minimizzazione dell'energia libera, con il sistema che alterna e combina le due in modo flessibile a seconda della struttura del gradiente di F.

Il collegamento con le teorie del controllo motorio è immediato. Il modello gerarchico di controllo motorio di Wolpert — che postula l'esistenza di modelli interni in avanti (forward models) e inversi (inverse models) usati per pianificare e correggere i movimenti — emerge naturalmente dalla struttura dell'aggiornamento variazionale: il forward model è il modello generativo $P(s|\psi, a)$, che predice le conseguenze sensoriali delle azioni; l'inverse model è implicito nella struttura del gradiente $\partial F/\partial a$, che inverte questa relazione per determinare quale azione riduce l'errore di predizione. Il principio di energia libera unifica questi due meccanismi in un unico framework variazionale.

IV.5 — L'inferenza attiva e i priori sulle azioni

Una raffinatezza importante del framework riguarda il modo in cui le preferenze — gli obiettivi dell'organismo — sono codificate. In un sistema di controllo tradizionale, gli obiettivi sono specificati esternamente come stati target da raggiungere. Nel framework del principio di energia libera, gli obiettivi sono codificati come *distribuzioni a priori* sugli stati sensoriali futuri: l'organismo agisce per portare la distribuzione dei segnali sensoriali futuri vicina alla distribuzione a priori che codifica le sue preferenze.

Formalmente, si introduce una distribuzione a priori $\tilde{P}(\tilde{s})$ sugli stati sensoriali generalizzati futuri — una "prior pragmatica" che assegna alta probabilità agli stati sensoriali associati a esiti desiderabili (basso dolore, adeguato apporto energetico, sicurezza). L'energia libera da minimizzare diviene allora:

$$G(a, \mu; \tilde{s}) = D_{KL}(Q(\tilde{\psi}; \mu) \,\|\, \tilde{P}(\tilde{\psi}, \tilde{s}))$$

dove $\tilde{P}(\tilde{\psi}, \tilde{s})$ è la distribuzione congiunta sugli stati nascosti e sensoriali futuri che incorpora sia il modello generativo che le preferenze pragmatiche. Questa quantità,

denominata *energia libera attesa* o *valore G* in analogia con il valore-G della teoria delle decisioni, si decompone in due termini:

$$G(a, \mu) = E_Q[-\ln \tilde{P}(\tilde{s})] + D_KL(Q(\tilde{\psi} \mid \tilde{s}; \mu) \parallel \tilde{P}(\tilde{\psi} \mid \tilde{s}))$$

Il primo termine è il valore pragmatico atteso negativo — misura quanto gli stati sensoriali futuri previsti si accordano con le preferenze. Il secondo termine è la riduzione attesa di incertezza — misura quanto le future osservazioni si aspettano di ridurre l'incertezza sugli stati del mondo. Questa decomposizione cattura la distinzione tra comportamento *estrinseco* (orientato al conseguimento di stati preferiti) e comportamento *epistemico* (orientato alla riduzione dell'incertezza, ovvero all'esplorazione informativa). La combinazione dei due è ciò che Friston chiama *inferenza attiva*: l'organismo non è né un puro sfruttatore delle sue conoscenze pregresse né un puro esploratore, ma un sistema che bilancia queste due spinte in modo variazionale.

La formalizzazione dell'esplorazione come termine dell'energia libera attesa è filosoficamente rilevante perché risolve in modo elegante il problema della curiosità: la curiosità non è un meccanismo ad hoc aggiunto al sistema di controllo, ma una conseguenza necessaria della minimizzazione dell'energia libera attesa in un sistema che tiene conto della propria incertezza futura. Un agente che minimizza G agirà in modo da ridurre la propria incertezza sugli stati del mondo quando questa riduzione è strumentale al raggiungimento di stati preferiti — e questa è precisamente la struttura del comportamento esplorativo adattivo.

IV.6 — L'equazione di Schrödinger stocastica e le fondazioni termodinamiche

Un risultato meno noto ma matematicamente affascinante che emerge dal principio di energia libera riguarda il legame con la meccanica quantistica e la termodinamica del non-equilibrio. Si consideri la distribuzione stazionaria $\varrho^*(x)$ di un sistema di Langevin:

$$dx = f(x)\, dt + \sqrt{(2D)}\, dW$$

La distribuzione stazionaria soddisfa l'equazione di Fokker-Planck stazionaria:

$$\nabla \cdot [\, f(x)\, \varrho^*(x) - D\, \nabla \varrho^*(x)\,] = 0$$

che può essere riscritta come:

$$\nabla \cdot [\, \varrho^*(x)\, (f(x) - D\, \nabla \ln \varrho^*(x))\,] = 0$$

Introducendo il potenziale di Gibbs $U(x) = -\ln \varrho^*(x)$ (a meno di costanti additive), si ottiene:

$$f(x) = -D \nabla U(x) + R(x)$$

dove R(x) è una corrente di conservazione solenoidale ($\nabla \cdot (\varrho^* R) = 0$) che cattura le componenti di non-equilibrio del flusso. Il flusso totale f si decompone dunque in un componente gradiente — che tende a portare il sistema verso i minimi dell'energia U — e una componente solenoidale — che produce moto ciclico attorno al bacino attrattore, violando il bilancio dettagliato e producendo entropia.

L'energia libera variazionale in questo contesto diviene:

$$F[\varrho] = \int \varrho(x) \, [\, U(x) + \ln \varrho(x) \,] \, dx = E_\varrho[U] - H[\varrho]$$

e la sua variazione temporale sotto la dinamica di Fokker-Planck è:

$$dF/dt = -\int \varrho(x) \, \| \, \nabla (U(x) + \ln \varrho(x)) \, \|^2 \, dx \leq 0$$

Questa disuguaglianza — la *disuguaglianza di H di Boltzmann* per i sistemi di diffusione — mostra che l'energia libera è una funzione di Lyapunov per la dinamica: diminuisce monotonicamente lungo le traiettorie finché non si raggiunge la distribuzione stazionaria. La connessione con il principio di energia libera di Friston è ora trasparente: il sistema biologico, operando con la sua struttura di Markov blanket, approssima questa discesa dell'energia libera attraverso l'aggiornamento delle rappresentazioni interne e l'esecuzione di azioni.

Il collegamento con la meccanica quantistica emerge attraverso la sostituzione $\varrho(x) = |\psi(x)|^2$ — la distribuzione di probabilità come modulo quadro di un'ampiezza complessa. Sotto questa sostituzione, l'equazione di Fokker-Planck per ϱ diviene, per sistemi con bilancio dettagliato, un'equazione di Schrödinger immaginaria:

$$\hbar \, \partial\psi/\partial\tau = -(\, \hbar^2/2m \,) \, \nabla^2\psi + V(x) \, \psi$$

dove $\tau = it$ è il tempo immaginario (rotazione di Wick). Questa corrispondenza — nota come dualità Fokker-Planck/Schrödinger — non è puramente formale: rivela che certi algoritmi di ottimizzazione variazionale nel caso classico hanno struttura analoga alla propagazione quantistica nel caso quantistico, aprendo la via a connessioni profonde tra la teoria dell'inferenza e la meccanica quantistica che rimangono un territorio parzialmente esplorato.

IV.7 — Critiche, limiti e prospettive aperte

Il principio di energia libera nella sua versione più ambiziosa è stato oggetto di critiche cogenti che meritano di essere affrontate con la stessa onestà intellettuale con cui si è esposto il formalismo. La più sostanziale riguarda la circolarità apparente della tesi. Se il principio afferma che qualsiasi sistema fisico che mantiene la propria integrità

minimizza l'energia libera, allora la tesi è vera per costruzione — è un'implicazione tautologica della struttura dei sistemi stocastici con distribuzione stazionaria — e non ha contenuto empirico aggiuntivo rispetto alla semplice affermazione che i sistemi viventi persistono. La risposta fristoniana è che il valore del principio non è empirico nel senso della falsificabilità popperiana, ma piuttosto normativo e computazionale: esso prescrive uno specifico modo di costruire modelli di sistemi biologici — attraverso la specifica del modello generativo — e quindi genera predizioni concrete una volta che il modello è specificato.

Una seconda critica riguarda la trattabilità computazionale dell'inferenza attiva per agenti con spazi di stati e di azioni ad altissima dimensionalità. Le derivazioni presentate nelle sezioni precedenti presuppongono che il sistema possa calcolare il gradiente dell'energia libera con precisione sufficiente — ma in sistemi nervosi reali, con milioni di neuroni e connessioni, la questione di quale algoritmo computazionale implementi effettivamente questo gradiente rimane aperta. Le proposte esistenti — dalle reti di neuroni a errore di predizione ai modelli di message passing corticale — sono suggestive ma non ancora ancorate a misurazioni neurofisiologiche sufficientemente precise da discriminare tra implementazioni alternative.

Rimane infine aperta la questione forse più profonda: il principio di energia libera offre un'ontologia o soltanto un'epistemologia dei sistemi cognitivi? Affermare che il sistema nervoso *minimizza* l'energia libera descrive la sua organizzazione funzionale, ma non risponde alla domanda di come questa minimizzazione si relazioni all'esperienza soggettiva — al *qualia* della percezione cosciente che costituisce il problema difficile della coscienza nel senso di Chalmers. Il framework variazionale, nella sua versione attuale, è radicalmente terza-persona: descrive la struttura computazionale dell'elaborazione percettiva dall'esterno, senza toccare la domanda di come e perché certi processi computazionali siano accompagnati da un'esperienza in prima persona. Questa lacuna non è un difetto peculiare del principio di energia libera — è condivisa da quasi tutta la scienza cognitiva computazionale — ma vale la pena tenerla presente come limite del programma, non come sua confutazione.

V. Codifica Predittiva Gerarchica: Implementazione Neurale del Gradiente Variazionale

---

V.1 — Il modello di Rao e Ballard: architettura e derivazione

Il lavoro di Rajesh Rao e Dana Ballard pubblicato su *Nature Neuroscience* nel 1999 costituisce il momento in cui il formalismo variazionale entra in contatto diretto con la neuroanatomia della corteccia visiva, producendo un modello che è al contempo computazionalmente derivato e neurobiologicamente plausibile. L'intuizione fondamentale — che le connessioni discendenti (feedback) nella gerarchia corticale

trasmettano predizioni mentre le connessioni ascendenti (feedforward) trasmettano errori di predizione — non era del tutto nuova: Mumford aveva anticipato questa idea nel 1992 in termini più informali, e la struttura gerarchica della corteccia visiva di macaco era stata ampiamente documentata dalla neuroanatomia di Felleman e Van Essen. Il contributo di Rao e Ballard fu di derivare questa architettura da principi variazionali di primo ordine e di mostrare che essa riproduce, senza parametri liberi addizionali, una serie di proprietà neurofisiologiche della corteccia visiva primaria — in particolare il fenomeno dell'adattamento neurale e gli effetti del contesto sul firing delle cellule semplici — che erano rimaste inspiegabili o richiedevano assunzioni ad hoc nei modelli precedenti.

Si consideri un modello generativo gerarchico a due livelli, dove si usa la notazione di Friston per renderlo compatibile con il framework generale della sezione II.6. Gli stati del primo livello $r^1 \in \mathbb{R}^k$ generano attivazioni sensoriali $s \in \mathbb{R}^m$ secondo:

$$s = f^0(r^1) + \varepsilon^0, \quad \varepsilon^0 \sim \mathcal{N}(0, (\Pi^0)^{-1})$$

Gli stati del secondo livello $r^2 \in \mathbb{R}^l$ generano le aspettative sugli stati del primo livello:

$$r^1 = f^1(r^2) + \varepsilon^1, \quad \varepsilon^1 \sim \mathcal{N}(0, (\Pi^1)^{-1})$$

dove $f^0 \colon \mathbb{R}^k \to \mathbb{R}^m$ e $f^1 \colon \mathbb{R}^l \to \mathbb{R}^k$ sono funzioni di mappatura nonlineari, e Π^0, Π^1 sono le matrici di precisione ai due livelli. L'energia libera variazionale per questo modello, sotto l'approssimazione di Laplace con stime puntuali $\hat{r}^1$, $\hat{r}^2$ degli stati nascosti, è:

$$F = (1/2)(s - f^0(\hat{r}^1))^T \Pi^0 (s - f^0(\hat{r}^1)) + (1/2)(\hat{r}^1 - f^1(\hat{r}^2))^T \Pi^1 (\hat{r}^1 - f^1(\hat{r}^2)) + \text{const}$$

Definendo gli errori di predizione ai due livelli:

$$\xi^0 = \Pi^0 (s - f^0(\hat{r}^1)) \in \mathbb{R}^m \quad \text{(errore sensoriale pesato per la precisione)}$$
$$\xi^1 = \Pi^1 (\hat{r}^1 - f^1(\hat{r}^2)) \in \mathbb{R}^k \quad \text{(errore del primo livello pesato)}$$

il gradiente dell'energia libera rispetto a $\hat{r}^1$ è:

$$\partial F / \partial \hat{r}^1 = -(\partial f^0 / \partial \hat{r}^1)^T \xi^0 + \xi^1$$

e il gradiente rispetto a $\hat{r}^2$ è:

$$\partial F / \partial \hat{r}^2 = -(\partial f^1 / \partial \hat{r}^2)^T \xi^1$$

Le equazioni di aggiornamento in tempo continuo per la discesa di gradiente sull'energia libera sono:

$$\dot{\hat{r}}^1 = \varkappa_1 \left[(\partial f^0 / \partial \hat{r}^1)^T \xi^0 - \xi^1 \right]$$

$$\dot{\hat{r}}^2 = \varkappa_2 \left[(\partial f^1 / \partial \hat{r}^2)^T \xi^1 \right]$$

dove $\varkappa_1, \varkappa_2 > 0$ sono costanti di velocità (learning rate). Queste equazioni hanno una lettura anatomica diretta. Il termine $(\partial f^0 / \partial \hat{r}^1)^T \xi^0$ è un segnale che sale dal livello sensoriale al primo livello nascosto: porta l'errore di predizione sensoriale, trasformato dal jacobiano transposto della funzione generativa — che corrisponde alle connessioni sinaptiche ascendenti dalla corteccia primaria alle aree di associazione. Il termine $-\xi^1$ è un segnale discendente che porta l'errore di predizione del livello superiore al livello inferiore — corrisponde alle connessioni di feedback dalla corteccia di ordine superiore verso la primaria. Il bilanciamento tra questi due segnali determina la stima di $\hat{r}^1$ che si stabilizza al punto fisso del sistema.

V.2 — Assegnazione neurale dei segnali e microarchitettura corticale

La domanda naturale è: quali popolazioni neuronali corticali codificano gli errori di predizione ξ e quali codificano le stime $\hat{r}$? Rao e Ballard proposero che i neuroni superficiali (strati II/III) della corteccia visiva codificassero gli errori di predizione, mentre i neuroni profondi (strati V/VI) codificassero le stime degli stati nascosti. Questa proposta era motivata dalla neuroanatomia: le connessioni feedforward che proiettano alle aree di ordine superiore originano prevalentemente dagli strati superficiali, mentre le connessioni di feedback che discendono dalle aree superiori verso quelle inferiori terminano prevalentemente negli strati profondi e nel primo strato. Se i segnali ascendenti portano errori di predizione e i segnali discendenti portano predizioni, la microarchitettura laminare della corteccia corrisponde esattamente alla struttura del circuito di codifica predittiva.

Una predizione verificabile di questo schema riguarda il comportamento degli errori di predizione in funzione della manipolazione delle aspettative. Se i neuroni superficiali codificano $\xi^0 = \Pi^0(s - f^0(\hat{r}^1))$, allora la loro risposta a uno stimolo sensoriale s dovrebbe essere attenuata — non zero, ma ridotta — quando lo stimolo è altamente previsto ($f^0(\hat{r}^1) \approx s$) e potenziata quando lo stimolo è inatteso ($f^0(\hat{r}^1) \ll s$ oppure $f^0(\hat{r}^1) \gg s$). Questo corrisponde al fenomeno della *soppressione del contesto* nella corteccia visiva primaria: la risposta di una cellula semplice V1 a un grating ottimale è attenuata quando il grating è circondato da un annulus dello stesso orientamento, che si assume generi aspettative concordanti attraverso connessioni orizzontali e di feedback. Analogamente, la risposta è potenziata quando lo stimolo è circondato da un orientamento ortogonale — un contesto che genera aspettative discordanti.

Friston e Kiebel (2009) e successivamente Bastos e collaboratori (2012) hanno esteso questo schema anatomico proposto da Rao e Ballard incorporando le oscillazioni

cerebrali. Misure di coerenza spettrale tra aree corticali mostrano che la comunicazione feedforward avviene prevalentemente nelle bande gamma (30-80 Hz), mentre la comunicazione feedback avviene prevalentemente nelle bande alpha-beta (8-30 Hz). Se le connessioni feedforward trasmettono errori di predizione e quelle feedback trasmettono predizioni, le diverse bande di frequenza potrebbero distinguere questi due tipi di segnale. Bastos e colleghi hanno proposto che i neuroni di errore di predizione (strati superficiali) oscillino preferenzialmente a frequenze gamma, mentre i neuroni che codificano le stime (strati profondi) oscillino nelle bande alpha-beta — e hanno trovato evidenza elettrofisiologica a supporto di questa dissociazione in registrazioni multi-area di macaco durante compiti di predizione visiva.

V.3 — Nonlinearità, modulazione e il ruolo della precisione

Il modello linearizzato di Rao e Ballard, pur catturando l'essenziale dell'architettura, è troppo semplice per rendere conto della ricchezza della fisiologia corticale reale. Le estensioni nonlineari richiedono di specificare la forma funzionale delle mappe f^0 e f^1. Una scelta naturale, motivata sia dai dati fisiologici che dalla teoria della rappresentazione, è di utilizzare funzioni di mistura gaussiana o, nel caso gerarchico profondo, composizioni di funzioni nonlineari — che è precisamente ciò che si ottiene nei modelli di reti neurali profonde.

Per le funzioni nonlineari, il jacobiano ($\partial f^1/\partial r^1$) che trasforma gli errori di predizione nelle aggiornamenti degli stati non è costante ma dipende dal punto di linearizzazione corrente. Questo significa che le "sinapsi efficaci" attraverso cui si propaga l'errore di predizione cambiano dinamicamente in funzione dello stato corrente del sistema — un meccanismo che potrebbe corrispondere alla modulazione nonlineare dell'efficacia sinaptica osservata fisiologicamente e che comprende meccanismi come la facilitazione sinaptica a breve termine, la depressione e la neuromodulazione.

Particolarmente rilevante è il ruolo delle matrici di precisione Π. Nella formulazione variazionale, la precisione è essa stessa un parametro da stimare: il sistema deve inferire non soltanto quale stato del mondo genera i propri segnali sensoriali, ma anche quanto ci si può fidare di questi segnali e delle proprie aspettative. L'aggiornamento variazionale della precisione al livello sensoriale richiede la minimizzazione di F rispetto a Π^0:

$$\partial F/\partial \Pi^0 = -(1/2)[\, \xi^0(\xi^0)^T/\Pi^0 - (\Pi^0)^{-1} \,]$$

Ponendo questo uguale a zero si ottiene la stima variazionale della precisione:

$$\Pi^0 = (\, E_Q[(s - f^0(\hat{r}^1))(s - f^0(\hat{r}^1))^T] \,)^{-1}$$

ovvero la matrice di precisione stimata è l'inversa della covarianza empirica dell'errore di predizione. Sistemi che aggiornano dinamicamente la propria stima di precisione possono dunque regolare il peso attribuito ai segnali sensoriali in funzione della loro

variabilità stimata — una modulazione top-down dell'efficacia sensoriale che costituisce la base formale dell'attenzione selettiva.

Il meccanismo è il seguente: un segnale con alta precisione stimata (Π^0 grande) produrrà errori di predizione $\xi^0 = \Pi^0(s - f^0(\hat{r}^1))$ più grandi per la stessa discrepanza sensoriale, e quindi aggiornamenti più rapidi delle stime degli stati nascosti. In effetti l'attenzione, in questa formulazione, non è un meccanismo di selezione che filtra i segnali sensoriali prima della loro elaborazione, ma una modulazione della precisione che determina quanto velocemente e quanto fortemente gli errori di predizione si propagano nella gerarchia. Questa concezione — l'attenzione come inferenza sulla precisione, o *precision-weighting* — è stata sviluppata dettagliatamente da Feldman e Friston (2010) e ha generato predizioni specifiche su come la modulazione attenzionale dovrebbe cambiare la risposta agli errori di predizione in V1 e nelle aree attentive di ordine superiore come V5 e il solco intraparietale.

V.4 — Apprendimento sinaptico come aggiornamento del modello generativo

Fino a questo punto si è discusso dell'inferenza percettiva — il processo rapido (da millisecondi a secondi) con cui il sistema aggiorna le proprie stime degli stati nascosti r^l dati i segnali sensoriali correnti. Ma il modello generativo stesso — le funzioni f^l e le strutture di precisione Π^l — deve essere acquisito attraverso l'esperienza. Questo apprendimento lento (da ore a anni) corrisponde, nel framework variazionale, alla minimizzazione dell'energia libera rispetto ai parametri del modello generativo $\theta = (f^0, f^1, \Pi^0, \Pi^1, ...)$.

Il gradiente dell'energia libera rispetto ai parametri sinaptici che definiscono f^l è:

$$\partial F / \partial \theta^l = -E_Q[\, \xi^l \cdot (\partial f^l / \partial \theta^l)^T \,]$$

che ha la forma di una regola di apprendimento hebbiano: l'aggiornamento sinaptico è proporzionale alla covarianza tra l'errore di predizione ξ^l e l'attività pre-sinaptica ($\partial f^l / \partial \theta^l$), mediata dalla distribuzione variazionale corrente Q. In forma discreta, per parametri sinaptici W^l che entrano linearmente nella funzione $f^l(r; W) = W^l r$, questo diviene:

$$\Delta W^l \propto E_Q[\, \xi^l \cdot (r^l)^T \,]$$

— il prodotto esterno tra errore di predizione post-sinaptico e attività pre-sinaptica, pesato per la precisione. Questa è precisamente la struttura della regola di Hebb modificata con segnale di errore — analoga alla regola delta nell'apprendimento supervisionato — ma derivata qui non come assunzione di apprendimento ma come conseguenza necessaria della minimizzazione dell'energia libera.

La connessione con la plasticità sinaptica biologica è suggestiva. La potenziation a lungo termine (LTP) e la depressione a lungo termine (LTD) sono i meccanismi

cellulari fondamentali della plasticità sinaptica Hebbiana. Recenti esperimenti hanno mostrato che l'LTP è più efficacemente indotta quando l'attività pre-sinaptica precede l'attività post-sinaptica in una finestra temporale di decine di millisecondi — la cosiddetta *spike-timing-dependent plasticity* (STDP). La derivazione variazionale suggerisce che la STDP sia la realizzazione biologica dell'aggiornamento dei pesi sinaptici nel senso dell'apprendimento variazionale: la finestra temporale asimmetrica della STDP riflette la struttura causale delle predizioni (pre-sinaptica) e degli errori di predizione (post-sinaptica) nel circuito di codifica predittiva.

V.5 — Codifica predittiva nei modelli profondi e connessione con le reti neurali artificiali

Il modello di Rao e Ballard a due livelli è, storicamente e concettualmente, il punto di partenza. Ma la corteccia visiva biologica comprende almeno sei aree visive gerarchicamente organizzate — V1, V2, V4, LOC, IT e PFC, ciascuna con molteplici sottoregioni — e vi è ampia evidenza che la complessità delle rappresentazioni aumenti sistematicamente lungo questa gerarchia, da caratteristiche locali di bordi e orientamenti in V1 fino a concetti astratti e categorie di oggetti nel lobo temporale inferiore. Il modello di codifica predittiva gerarchica profonda è dunque naturalmente:

$$
\begin{aligned}
s &= f^{0}(r^{1}) + \varepsilon^{0} \\
r^{1} &= f^{1}(r^{2}) + \varepsilon^{1} \\
r^{2} &= f^{2}(r^{3}) + \varepsilon^{2} \\
&\vdots \\
r^{L-1} &= f^{L-1}(r^{L}) + \varepsilon^{L-1}
\end{aligned}
$$

con L livelli gerarchici, e l'energia libera complessiva è la somma degli errori di predizione pesati per le precisioni a ogni livello:

$$F = \sum_{l=0}^{L-1} \frac{1}{2} (\varepsilon^{l})^{T} \Pi^{l} \varepsilon^{l} + \text{const}$$

Le equazioni di aggiornamento in forma vettoriale compatta diventano, per ogni livello interno $l \in \{1, \ldots, L-1\}$:

$$\dot{\hat{r}}^{l} = \kappa^{l} \left[\left(\partial f^{l-1} / \partial \hat{r}^{l}\right)^{T} \xi^{l-1} - \xi^{l} \right]$$

dove $\xi^{l-1} = \Pi^{l-1}(\hat{r}^{l-1} - f^{l-1}(\hat{r}^{l}))$ è l'errore di predizione che il livello l deve spiegare (proveniente dal livello inferiore) e $\xi^{l} = \Pi^{l}(\hat{r}^{l} - f^{l}(\hat{r}^{l+1}))$ è l'errore di predizione generato dal livello l rispetto al livello superiore. Il livello l funge dunque da mediatore: deve essere sufficientemente ricco da spiegare ciò che sale dal basso mentre rispettando i vincoli imposti dall'alto.

Questa struttura è sorprendentemente simile — non identica — all'architettura delle reti neurali profonde con connessioni skip e alle reti ResNet. Nelle ResNet, ogni blocco residuale calcola:

$$y = x + \mathcal{F}(x, \{W_i\})$$

dove x è l'input del blocco e $\mathcal{F}(x, \{W_i\})$ è la "funzione residuale" appresa. La corrispondenza con la codifica predittiva è che $\mathcal{F}$ rappresenta l'errore di predizione — la discrepanza tra la rappresentazione al livello corrente e quella predetta dal livello precedente — e la connessione skip corrisponde alla trasmissione della predizione discendente. Questa non è una coincidenza superficiale: Lotter, Kreiman e Cox (2017) hanno costruito esplicitamente architetture di rete neurale (PredNet) ispirate alla codifica predittiva gerarchica, e hanno mostrato che queste reti apprendono rappresentazioni video più efficienti e biologicamente più plausibili rispetto alle architetture standard, con la proprietà notevole di estrarre automaticamente rappresentazioni di movimento e struttura 3D senza supervisione esplicita.

VI. Implicazioni per la Filosofia della Mente: Realismo, Enattivismo e il Problema Difficile

---

VI.1 — Realismo strutturale e il modello interno del mondo

Il formalismo variazionale sviluppato nelle sezioni precedenti ha conseguenze non banali per il dibattito tra le diverse versioni del realismo in filosofia della scienza e della mente. La posizione più immediata che emerge dal framework — la percezione come inferenza alla spiegazione migliore, la rappresentazione percettiva come ipotesi a posteriori — potrebbe sembrare iscriversi naturalmente nell'antirealismo epistemico: se ciò che percepiamo è sempre e soltanto la distribuzione a posteriori $P(\psi|s)$ sul modello interno, mai il mondo ψ direttamente, allora non abbiamo accesso epistemico alle cause profonde del segnale sensoriale, e la veridicità delle nostre rappresentazioni non è garantibile a priori.

Questa conclusione, tuttavia, è precipitosa. Il framework variazionale non implica l'antirealismo, ma piuttosto una versione sofisticata di *realismo strutturale* nel senso di John Worrall e James Ladyman. La tesi del realismo strutturale è che ciò che le teorie scientifiche — e, nell'estensione che qui si propone, i modelli percettivi — colgono veridicamente del mondo non sono gli oggetti particolari nelle loro proprietà intrinseche, bensì le relazioni strutturali tra entità che, nella fase pre-teorica, vengono denominate "oggetti". Il modello generativo $P(s|\psi)$ che il sistema nervoso ha interiorizzato attraverso l'adattamento filogenetico e l'apprendimento ontogenetico non garantisce che le variabili ψ corrispondano a oggetti con proprietà intrinseche determinate, ma garantisce — nella misura in cui l'apprendimento ha successo — che le relazioni strutturali catturate dal modello si specchino nelle relazioni strutturali del mondo fisico che ha generato i segnali sensoriali durante l'apprendimento.

Il punto decisivo è il seguente. L'apprendimento variazionale non è casuale: esso minimizza l'energia libera, ovvero massimizza la log-evidenza del modello $P(s)$. Un

modello che cattura le strutture causali reali del mondo produce una log-evidenza più alta — predice meglio i segnali sensoriali futuri — rispetto a un modello che non le cattura. La pressione selettiva dell'adattamento biologico su scale evolutive, unita alla plasticità sinaptica nell'individuo, produce dunque un progressivo affinamento del modello generativo verso strutture che si accordano con quelle del mondo. Non c'è nessuna garanzia metafisica che questo processo converga verso una descrizione completa e veritiera della realtà in sé; ma c'è una garanzia statistica — mediata dai teoremi di consistenza dell'inferenza bayesiana — che, con sufficiente esperienza in un ambiente stazionario, il modello converga verso la distribuzione generativa vera da cui sono stati campionati i dati. Il realismo strutturale emerso dal processo variazionale è dunque una forma di realismo pragmatico-statistico: vero non di necessità logica ma di convergenza statistica.

VI.2 — Enattivismo computazionale: percezione come azione di campionamento

Una tensione profonda attraversa il dibattito tra il framework della codifica predittiva e le tradizioni enattiviste in filosofia della mente. L'enattivismo, nella formulazione di Maturana e Varela e nella sua variante più filosoficamente elaborata di Alva Noë, sostiene che la percezione non è una rappresentazione interna del mondo costruita dal sistema nervoso, ma un processo dinamico di accoppiamento strutturale tra l'organismo e il suo ambiente — un'attività, non un risultato. La cognizione è *embodied* e *enacted*: dipende essenzialmente dalla struttura corporea dell'organismo e dai pattern di interazione senso-motorii che essa rende possibili.

A prima vista, il framework variazionale sembra contrapposto all'enattivismo: è fondato sul concetto di modello interno, di rappresentazione, di distribuzione a posteriori — tutti concetti che l'enattivismo radicale rifiuta come residui del paradigma computazionalista cartesiano. Tuttavia, una lettura più attenta rivela una convergenza sostanziale che merita di essere articolata con precisione.

Il principio di energia libera nella sua versione estesa — quella che include l'azione come minimizzazione dell'energia libera attraverso la modifica dell'ambiente sensoriale — non è compatibile con una concezione della percezione come processo puramente passivo e interiore. L'organismo agisce per portare i propri segnali sensoriali in accordo con le proprie predizioni, e queste azioni plasmano l'ambiente che a sua volta genera i segnali futuri. La percezione e l'azione sono intrecciate in un ciclo continuo in cui il confine tra "dentro" e "fuori" è funzionale, non metafisico: il Markov blanket definisce questo confine operativamente, come superficie attraverso cui avviene l'accoppiamento tra stati interni ed esterni, non come barriera che separa un soggetto da un mondo indipendente.

In questa lettura, ciò che si propone è un *enattivismo computazionale*: una posizione che prende sul serio l'insistenza enattivista sull'accoppiamento senso-motorio e sull'embodiment, ma lo inquadra all'interno del formalismo variazionale invece di opporsi ad esso. Il modello generativo non è una rappresentazione cartografica del mondo esterno conservata in un teatro interno; è una struttura che codifica le

contingenze senso-motorie — i pattern di "se faccio questo, mi aspetto quello sensorialmente" — che costituiscono le abilità percettive dell'organismo. Questa idea, sviluppata da Jakob Hohwy in parziale dialogo con l'enattivismo di Noë, riceve ora una formalizzazione precisa: il modello generativo $P(s|\psi, a)$ che include gli stati attivi a come variabili del modello è precisamente una codifica delle contingenze senso-motorie nel senso enattivista. Apprendere a percepire significa apprendere le dipendenze tra azioni e conseguenze sensoriali — ovvero costruire un modello generativo che si estende oltre la frontiera del Markov blanket in entrambe le direzioni.

VI.3 — La struttura intensionale della rappresentazione variazionale

Un'ulteriore implicazione filosofica del framework riguarda la struttura delle rappresentazioni mentali e il loro rapporto con la tradizione della logica intensionale e della semantica dei mondi possibili. In filosofia del linguaggio e della mente, si distingue tra la *estensione* di un concetto — l'insieme degli oggetti a cui si applica — e la sua *intensione* — il modo in cui il concetto "coglie" o "presenta" quegli oggetti, che determina il suo contributo al contenuto dei giudizi in cui compare. Due concetti possono avere la stessa estensione — "creature con cuore" e "creature con rene" si applicano agli stessi organismi nel mondo attuale — pur avendo intensioni diverse e contribuendo diversamente al contenuto dei giudizi.

La distribuzione a posteriori $P(\psi|s)$ del sistema percettivo ha una struttura naturalmente intensionale. Essa non seleziona un singolo stato del mondo (l'estensione) ma distribuisce credenze su un insieme di stati possibili, ponderati per la loro plausibilità a posteriori. Questo spazio di stati possibili è esattamente lo spazio dei "mondi possibili" della semantica modale, e la distribuzione a posteriori è formalmente analoga a una misura di probabilità sui mondi possibili del tipo usato da David Lewis nella semantica dei condizionali. La credenza percettiva "vedo un cubo rosso" non afferma l'esistenza di un cubo rosso (estensione) ma piuttosto assegna alta probabilità agli stati del mondo che contengono un cubo rosso nella posizione percepita (intensione probabilistica). Le illusioni percettive sono esattamente i casi in cui l'intensione — la distribuzione a posteriori — diverge dall'estensione — lo stato del mondo effettivo.

Questo legame tra framework variazionale e semantica intensionale ha un'applicazione concreta nel problema della *opacità referenziale*. Le credenze percettive sono intensionali — non permettono la sostituzione salva veritate di termini co-referenziali — e questa opacità riflettte la struttura della distribuzione a posteriori: il sistema non ha accesso all'identità degli oggetti nel mondo, ma soltanto alla struttura probabilistica del proprio modello generativo. Se il sistema conosce che "lo stimolo nella posizione P" causa segnali s con alta probabilità, sa altresì che "il cubo rosso in posizione P" causa gli stessi segnali, ma non può "sapere" che i due descrittori si riferiscono allo stesso oggetto a meno che il suo modello generativo non contenga esplicitamente questa informazione strutturale. L'opacità referenziale è dunque, in questo schema, non un difetto logico delle credenze mentali ma un riflesso della struttura parziale del modello generativo.

VI.4 — Il problema difficile della coscienza nel quadro variazionale

La questione più profonda — e più refrattaria — che il framework variazionale deve affrontare è quella del *problema difficile della coscienza* nel senso di David Chalmers. Chalmers distingue i "problemi facili" della coscienza — spiegare come il cervello integra l'informazione, discrimina gli stimoli, controlla il comportamento — dai "problemi difficili" — spiegare perché e come certi processi fisici siano accompagnati da esperienza soggettiva, dal "qualcosa che è come essere" nel senso di Thomas Nagel. I problemi facili sono difficili ma plausibilmente trattabili con metodi computazionali e neuroscientifici; il problema difficile sembra resistere in linea di principio a ogni spiegazione funzionale.

Il framework variazionale risolve, almeno in linea di principio, i problemi facili con notevole eleganza. La discriminazione sensoriale emerge dalla struttura dell'errore di predizione; l'integrazione dell'informazione corrisponde alla costruzione di stati nascosti condivisi ai livelli gerarchici superiori; il controllo del comportamento è la dinamica dell'inferenza attiva. Ma il problema difficile rimane insoluto — e qui è necessario essere onesti riguardo ai limiti del formalismo.

Esistono due strategie principali con cui i proponenti del framework variazionale hanno affrontato questa questione. La prima è la *deflationary strategy*: sostenere che il problema difficile è mal posto, che la distinzione tra "problemi facili" e "problema difficile" è illusoria, e che spiegare i meccanismi funzionali della coscienza equivale a spiegare tutto ciò che è genuinamente spiegabile sulla coscienza. Il residuo — il "qualcosa che è come essere" — non è una questione empirica ma una confusione concettuale che si dissolve con un'analisi più accurata. Friston ha simpatizzato parzialmente con questa posizione, suggerendo che l'esperienza soggettiva emerga dalla prospettiva interna della minimizzazione dell'energia libera — il "feel" della percezione è il sistema che si vede dall'interno mentre compie inferenza — ma senza sviluppare questa idea in un argomento formalmente preciso.

La seconda strategia è il *panpsichismo strutturale*: sostenere che le proprietà fenomeniche sono costitutive della struttura causale del mondo a tutti i livelli di organizzazione, e che l'esperienza cosciente complessa emerge da combinazioni di proprietà fenomeniche elementari in sistemi sufficientemente organizzati. In questa prospettiva, il framework variazionale descrive i correlati strutturali dell'esperienza — le relazioni matematiche tra stati variazionali — mentre le proprietà intrinseche degli stati rimangono fenomeniche. Questa posizione, sostenuta tra gli altri da Phillip Goff in forme non direttamente connesse al framework variazionale, incontra il proprio problema serio nel cosiddetto *combination problem*: come si combinano le proprietà fenomeniche elementari per produrre l'esperienza unitaria e ricca di un organismo complesso? La risposta variazionale — che l'integrazione dell'informazione attraverso la gerarchia corticale costituisce l'integrazione dell'esperienza — è suggestiva ma circolare finché non si specifica il meccanismo di questa "integrazione fenomenologica".

Una terza via, meno battuta ma filosoficamente interessante, emerge dal legame tra il framework variazionale e la teoria dell'informazione integrata (IIT) di Giulio Tononi. L'IIT propone che la coscienza sia identica all'informazione integrata Φ — una misura della quantità di informazione generata dal sistema come un tutto al di là di quella generata dalle sue parti. Un sistema ha esperienza cosciente se e soltanto se il suo Φ è positivo, e la ricchezza dell'esperienza è proporzionale al suo Φ. Ora, la struttura del modello generativo gerarchico della codifica predittiva produce naturalmente sistemi con Φ positivo elevato: le rappresentazioni agli strati superiori della gerarchia integrano l'informazione proveniente da domini sensoriali multipli in modi che non sono riducibili alle rappresentazioni dei singoli livelli inferiori. Se l'IIT è corretta, i sistemi di codifica predittiva gerarchica sarebbero coscienti in misura proporzionale alla loro profondità e densità di integrazione — una predizione che sarebbe, almeno teoricamente, verificabile.

Tuttavia, è necessario riconoscere che nessuna di queste strategie risolve il problema difficile nel senso di Chalmers. Il problema difficile non è una questione empirica che si risolve accumulando dati neurofisiologici o raffinando il formalismo variazionale: è una questione metafisica che riguarda la relazione tra il dominio della descrizione fisica — inclusa quella variazionale — e il dominio dell'esperienza in prima persona. Il formalismo variazionale può descrivere con crescente precisione i meccanismi computazionali e neurali della percezione; non può, per sua natura, spiegare perché questi meccanismi siano accompagnati da qualcosa che "è come" percepire. Riconoscere questo limite non è una sconfitta del programma variazionale: è la condizione di onestà intellettuale che gli consente di mantenere il proprio rigore dove esso è pertinente, senza pretendere di risolvere ciò che si trova al di là del suo orizzonte naturale.

VI.5 — Psicopatologia come inferenza patologica: schizofrenia e autismo

Una delle applicazioni più fruttuose e concretamente verificabili del framework variazionale riguarda la comprensione di alcune condizioni psicopatologiche come disruzioni specifiche del processo di inferenza. Se la percezione normale è il risultato di un equilibrio fine tra errori di predizione sensoriali e aspettative gerarchiche, pesate per le rispettive precisioni, allora alterazioni in questo equilibrio producono disturbi percettivi caratteristici — e la firma matematica di queste alterazioni consente predizioni quantitative testabili.

La schizofrenia, in questa prospettiva, può essere interpretata come una condizione di *aberrant salience* — salienzaa aberrante degli errori di predizione — causata da un'eccessiva attribuzione di precisione ai segnali sensoriali rispetto alle aspettative gerarchiche. Formalmente, se la matrice di precisione sensoriale Π^0 è sopravvalutata rispetto alle precisioni dei livelli superiori Π^l ($l > 0$), gli errori di predizione sensoriali $\xi^0 = \Pi^0(s - f^0(\hat{r}^1))$ avranno un peso eccessivo nelle equazioni di aggiornamento, impedendo alle aspettative di alto livello di sopprimere appropriatamente gli errori di basso livello. Il risultato è che segnali sensoriali casuali o privi di significato strutturale vengono trattati come portatori di informazione causale rilevante — la base

computazionale delle idee di riferimento e del pensiero magico. Le allucinazioni uditive, in questa lettura, emergono quando la precisione degli stati nascosti interni al livello linguistico supera quella dei segnali sensoriali esterni — il pensiero interno viene "percepito" come proveniente dall'esterno perché il suo errore di predizione è attribuito a una causa sensoriale esterna anziché a una fonte interna.

Questa analisi genera predizioni verificabili: i pazienti schizofrenici dovrebbero mostrare risposte potenziate agli stimoli inattesi (violazioni delle aspettative) e risposte ridotte alla soppressione del contesto rispetto ai controlli sani. L'evidenza disponibile — comprese misurazioni di potenziali evocati (ERP) come la negatività da mancata corrispondenza, MMN — è in ampio accordo con questa predizione: la MMN, che misura la risposta neurale agli stimoli acustici devianti in una sequenza di standard, è sistematicamente ridotta nella schizofrenia, indicando una capacità ridotta di utilizzare le aspettative per sopprimere le risposte agli stimoli standard e amplificare quelle agli stimoli inattesi.

Il disturbo dello spettro autistico suggerisce, dal framework variazionale, un'alterazione di segno opposto o complementare. La proposta di Pellicano e Burr (2012), ripresa e formalizzata da Lawson, Rees e Friston (2014), è che l'autismo sia caratterizzato da priori iperpredicibili ma poco flessibili — un sistema che assegna a priori stretti e difficilmente aggiornabili sugli stati del mondo, producendo rigidità percettiva e comportamentale. In termini formali, le distribuzioni a priori $P(\psi)$ avrebbero varianza molto ridotta (alta precisione a priori), il che implicherebbe errori di predizione sistematicamente grandi quando l'ambiente presenta variabilità inaspettata — esattamente la sensibilità all'imprevedibilità che caratterizza molti individui con autismo. La notevole acuità percettiva in certi domini (l'iper-percezione dei dettagli sensoriali) sarebbe la controparte positiva dello stesso meccanismo: quando le aspettative di alto livello non sopprimono efficacemente i dettagli sensoriali di basso livello, questi ultimi diventano accessibili alla coscienza con una ricchezza inusuale.

La verifica formale di queste ipotesi richiede metodi di inferenza bayesiana computazionale applicati ai dati comportamentali e neurofisiologici — un programma di *neuroscienze computazionali della psicopatologia* che è attualmente in sviluppo attivo nei laboratori di Friston, di Chris and Uta Frith, e di numerosi gruppi a loro affiliati. Il valore di questo programma non risiede soltanto nella sua capacità predittiva, ma nella coerenza del framework che lo sostiene: le ipotesi sulla psicopatologia non sono proposte ad hoc ma derivate dalla stessa struttura matematica che descrive la percezione normale, con alterazioni parametriche specifiche e quantificate.

VI.6 — Verso una sintesi: il continuum dell'inferenza variazionale

Il percorso che questo saggio ha intrapreso — dalle radici epistemologiche helmholtziane alla geometria dell'informazione di Amari, dal principio di energia libera di Friston alle implicazioni per il problema difficile della coscienza — converge verso una visione che potrebbe essere caratterizzata come *naturalismo epistemico variazionale*: la tesi che i processi cognitivi fondamentali siano istanze di inferenza variazionale, che

questa inferenza sia implementata in sistemi neurali gerarchici, e che l'ontologia della mente — inclusi i concetti di credenza, rappresentazione, percezione e azione — riceva la sua articolazione più precisa all'interno di questo framework formale.

Questa visione non è esente da problemi aperti, né pretende di esserlo. Il problema della trattabilità computazionale dell'inferenza profonda, la questione di quale famiglia variazionale sia effettivamente implementata nel sistema nervoso, il problema della regolarizzazione e della generalizzazione nei modelli gerarchici profondi, e soprattutto il problema difficile della coscienza — tutte queste questioni restano parzialmente o interamente aperte. Ma l'apertura di queste questioni non è un segno di debolezza del framework; è il segno della sua maturità come programma di ricerca nel senso di Imre Lakatos: un nucleo duro di principi matematicamente costrittivi, un'euristica positiva che genera predizioni verificabili, e un'euristica negativa che delimita il dominio di applicazione senza ridurlo artificialmente.

Ciò che il framework variazionale ha compiuto — e che nessuna delle teorie della percezione precedenti aveva compiuto con la stessa sistematicità — è di rendere inseparabili e mutuamente illuminanti quattro ordini di considerazioni che la tradizione aveva mantenuto distinti: il problema normativo (quale è l'inferenza razionale ottimale?), il problema computazionale (quale algoritmo può approssimarla?), il problema implementativo (come è realizzato nei circuiti neurali?), e il problema filosofico (che tipo di conoscenza del mondo produce, e come si relaziona all'esperienza soggettiva?). Il formalismo variazionale risponde in modo integrato a questi quattro livelli — non perfettamente, non definitivamente, ma con un grado di coerenza e fertilità che lo rende, allo stato attuale della scienza cognitiva, il candidato più promettente a una teoria unificata della mente.

Naturalismo Epistemico Variazionale: Una Sintesi Provvisoria

La coerenza di un programma

Un saggio che attraversa la fisiologia della percezione di Helmholtz, il formalismo variazionale dell'inferenza bayesiana, la geometria riemanniana di Amari, il principio di energia libera di Friston e il problema difficile della coscienza di Chalmers rischia di apparire come un catalogo di temi accostati per affinità tematica piuttosto che come un'argomentazione unitaria. Vale dunque la pena esplicitare, in sede conclusiva, la tesi che ha attraversato tutti i movimenti precedenti e che conferisce al percorso la sua coerenza interna.

La tesi è questa: la percezione è inferenza variazionale, e questa affermazione non è una metafora né un programma di ricerca da realizzare in futuro, ma una descrizione matematicamente precisa di ciò che il sistema nervoso compie quando traduce flussi di segnali fisici in rappresentazioni del mondo. Ogni termine di questa affermazione porta il peso di una derivazione o di un argomento specifico. "Inferenza" perché il problema dell'ottica inversa è genuinamente mal posto nel senso di Hadamard, e qualsiasi soluzione stabile richiede l'introduzione di conoscenza a priori — non come scelta epistemica ma come necessità matematica. "Variazionale" perché la distribuzione a posteriori esatta è computazionalmente intrattabile per qualsiasi modello generativo non banale, e l'approssimazione mediante ottimizzazione dell'ELBO è la sostituzione computazionalmente trattabile più principiata disponibile. "Il sistema nervoso compie" perché l'architettura laminare della corteccia, la distribuzione delle connessioni feedforward e feedback, la modulazione della precisione attraverso i neurotrasmettitori, e la dinamica delle oscillazioni corticali sono tutti coerenti con — e in molti casi predetti da — la struttura delle equazioni variazionali, non semplicemente compatibili con esse.

Questa tesi ha quattro corollari che il saggio ha sviluppato progressivamente.

Il primo è che la geometria dello spazio delle distribuzioni — la varietà riemanniana con metrica di Fisher, le connessioni duali di Amari, le geodetiche esponenziali e di mistura — non è uno strumento matematico esterno applicato al problema della percezione, ma la struttura intrinseca dello spazio entro cui l'aggiornamento percettivo si svolge. La discesa di gradiente naturale, che usa la metrica di Fisher anziché quella euclidea, non è un'ottimizzazione migliore di quella euclidea nello stesso spazio: è l'ottimizzazione nell'unico spazio geometricamente canonico per i problemi di inferenza, quello in cui le distanze hanno significato statistico invariante rispetto alla parametrizzazione. Che le equazioni di codifica predittiva pesino gli errori con le matrici di precisione è, in questa luce, la firma della discesa di gradiente naturale — non una scelta di modellazione, ma la conseguenza necessaria della struttura geometrica del problema.

Il secondo corollario è che l'azione non è separabile dalla percezione se non analiticamente, non ontologicamente. Il principio di energia libera esteso mostra che percezione e azione sono due modi di minimizzare la stessa funzione — l'energia libera variazionale — attraverso due leve diverse: aggiornare le rappresentazioni interne tenendo fisso l'ambiente, oppure modificare l'ambiente tenendo fisse le rappresentazioni. Questa simmetria formale dissolve il dualismo percettivo-motorio che ha attraversato la psicologia cognitiva per decenni, non per decreto filosofico ma per struttura matematica. L'enattivismo, che aveva intuito questa simmetria in termini fenomenologici, riceve qui la sua formalizzazione quantitativa.

Il terzo corollario riguarda l'apprendimento. L'aggiornamento sinaptico che il sistema nervoso compie su scale temporali lunghe non è un processo separato dall'inferenza percettiva: è l'inferenza percettiva applicata ai parametri del modello generativo anziché agli stati nascosti. La stessa struttura di gradiente — errore di predizione pesato per la precisione, correlato con l'attività pre-sinaptica — governa entrambi i

processi a scale temporali diverse. La plasticità sinaptica biologica, nella misura in cui segue la spike-timing-dependent plasticity, è la realizzazione neurale di questo aggiornamento variazionale lento. Il sistema nervoso non impara prima e poi percepisce: apprende percependo, aggiornando simultaneamente le stime degli stati correnti e i parametri del modello che li genera.

Il quarto corollario, il più delicato, riguarda i limiti del framework. L'inferenza variazionale può rendere conto, con precisione crescente, di tutto ciò che la scienza cognitiva chiama "problemi facili" della coscienza: discriminazione, integrazione, controllo, attenzione, apprendimento, psicopatologia. Non può — e non è chiaro che possa in linea di principio — rendere conto del fatto che questi processi siano accompagnati da esperienza soggettiva. Questo limite non è un'omissione colmabile con ulteriore elaborazione tecnica: è il confine del dominio in cui la descrizione in terza persona, per quanto matematicamente ricca, resta descrizione in terza persona. Il problema difficile della coscienza non si risolve affinando il formalismo variazionale; si affronta, con gli strumenti della filosofia, riconoscendo che la struttura matematica descrive i correlati dell'esperienza senza esaurirne il contenuto fenomenico.

La fecondità di un'incompletezza

Ciò che rende il programma variazionale intellettualmente vitale non è la sua completezza — che è lungi dall'essere raggiunta — ma la sua fecondità: la capacità di generare domande nuove e precise in domini che sembravano separati. La connessione tra la geometria di Wasserstein e il rumore neuronale funzionale, tra la dualità Fokker-Planck/Schrödinger e i possibili fondamenti quantistici dell'inferenza neurale, tra la struttura dell'energia libera attesa e la curiosità epistemica come proprietà emergente dell'ottimizzazione variazionale — queste non sono speculazioni a margine del programma, ma direzioni di ricerca che emergono necessariamente dalla struttura interna del formalismo.

Vi è infine una considerazione di carattere epistemologico generale che merita di chiudere questo percorso. Il formalismo variazionale, nella misura in cui descrive il modo in cui il cervello costruisce la propria rappresentazione del mondo, è al tempo stesso una teoria del cervello e una teoria della conoscenza — una neuroscienza e un'epistemologia naturale nel senso di Quine. Ma a differenza dell'epistemologia naturalizzata quiniana, che si limitava a ricondurre la teoria della conoscenza alla psicologia cognitiva, il framework variazionale offre un apparato normativo preciso: l'inferenza bayesiana è la teoria della conoscenza razionale ottimale, e il sistema nervoso la approssima attraverso l'inferenza variazionale. La norma e il fatto sono qui intrecciati in modo non circolare: la norma è derivata da principi di coerenza probabilistica indipendentemente dalla biologia, e il fatto è che il sistema nervoso, entro i suoi vincoli computazionali, si avvicina a questa norma in modo misurabile.

Questa posizione — chiamiamola *naturalismo epistemico variazionale* — non riduce l'epistemologia alla neuroscienze né le neuroscienze all'epistemologia. Sostiene invece che le due discipline convergono verso la stessa struttura matematica da direzioni diverse: la prima partendo dalla domanda normativa su quale sia la credenza razionale,

la seconda partendo dalla domanda descrittiva su come il cervello elabori l'informazione. Quando una struttura matematica appare necessaria da entrambe le direzioni, questo è di solito un segnale che si è toccato qualcosa di profondo — non una verità ultima, ma una articolazione genuina della realtà piuttosto che un artefatto della prospettiva adottata. Il formalismo variazionale, nella sua attuale maturità, ha la forma di una tale articolazione. Resta da vedere fino a dove essa si estende.

Appendice Tecnica — Dall'Equazione di Langevin alla Fokker-Planck: Derivazione Completa e Dualità con l'Equazione di Schrödinger

---

A.1 — Motivazione e contesto

Nel corpo principale del saggio, in particolare nella sezione IV.6, si è affermato che la discesa di gradiente dell'energia libera variazionale in senso di Wasserstein produce la dinamica di Fokker-Planck, e che quest'ultima — mediante rotazione di Wick al tempo immaginario — è formalmente duale a un'equazione di Schrödinger. Entrambe queste affermazioni sono state enunciate senza derivazione, per non interrompere il filo argomentativo principale. Questa appendice colma quella lacuna, fornendo le derivazioni complete con tutti i passaggi esplicitati.

Le tre sezioni che seguono coprono, nell'ordine: la derivazione dell'equazione di Fokker-Planck a partire dall'equazione di Langevin stocastica, mediante il metodo della funzione caratteristica; la derivazione dell'equazione stazionaria di Fokker-Planck e la caratterizzazione della distribuzione stazionaria; la dualità con l'equazione di Schrödinger in tempo immaginario mediante la trasformazione di Doob e la rotazione di Wick.

A.2 — Dall'equazione di Langevin all'equazione di Fokker-Planck

A.2.1 — L'equazione di Langevin generalizzata

Si considera un sistema dinamico stocastico descritto da uno stato $x(t) \in \mathbb{R}^n$ che evolve secondo l'equazione differenziale stocastica di Itô:

$$dx = f(x, t)\, dt + \sigma(x, t)\, dW(t) \qquad (A.1)$$

dove:

$f: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ è il drift deterministico (campo vettoriale)
$\sigma: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{n \times m}$ è la matrice di diffusione
$W(t) \in \mathbb{R}^m$ è un processo di Wiener standard (moto browniano)

Il processo di Wiener W(t) soddisfa:

$$W(0) = 0$$
$$E[dW(t)] = 0$$
$$E[dW(t)\, dW(t)^T] = I\, dt \qquad \text{(A.2)}$$

dove I è la matrice identità m×m. La matrice di diffusione $D(x,t) = \sigma(x,t)\, \sigma(x,t)^T$ è una matrice n×n simmetrica semidefinita positiva.

L'interpretazione fisica dell'equazione (A.1) è quella di un sistema soggetto a una forza deterministica f(x,t) e a fluttuazioni stocastiche gaussiane di covarianza D(x,t) dt per passo temporale infinitesimo.

A.2.2 — La densità di transizione e l'espansione di Kramers-Moyal

Sia $p(x, t \mid x_0, t_0)$ la densità di transizione del processo — la densità di probabilità di trovarsi nello stato x al tempo t, dato lo stato iniziale x_0 al tempo t_0. L'obiettivo è derivare l'equazione differenziale alle derivate parziali che governa l'evoluzione temporale di p.

Si considera la funzione caratteristica del processo nell'intervallo [t, t + dt]. Per un'osservabile arbitraria $\varphi(x) \in C^2(\mathbb{R}^n)$, si calcola la variazione attesa in dt usando la formula di Itô. Per una funzione sufficientemente regolare $\varphi: \mathbb{R}^n \to \mathbb{R}$, la formula di Itô afferma:

$$d\varphi(x) = (\partial\varphi/\partial x_i)\, dx_i + (1/2)(\partial^2\varphi/\partial x_i \partial x)\, dx_i\, dx \qquad \text{(A.3)}$$

(somma sugli indici ripetuti). Sostituendo (A.1) nell'espressione (A.3) e usando le regole del calcolo stocastico di Itô:

$$dt \cdot dt = 0$$
$$dt \cdot dW = 0$$
$$dW \cdot dW = \delta\, dt$$

si ottiene:

$$d\varphi(x) = (\partial\varphi/\partial x_i)\, f_i(x,t)\, dt + (\partial\varphi/\partial x_i)\, \sigma_i\square(x,t)\, dW + (1/2)(\partial^2\varphi/\partial x_i \partial x)\, \sigma_i\square\, \sigma\, dt \qquad \text{(A.4)}$$

Prendendo il valore atteso — il termine con dW ha valore atteso nullo per la proprietà del processo di Wiener — si ha:

$$E[d\varphi/dt] = E[\, f_i(x,t)\, \partial\varphi/\partial x_i + (1/2)\, D_{i\square}(x,t)\, \partial^2\varphi/\partial x_i \partial x \,] \quad (A.5)$$

dove $D_{i\square} = \sigma_{i\square}\, \sigma$ è la componente (i,j) della matrice di diffusione. Scrivendo esplicitamente il valore atteso come integrale rispetto alla densità p(x,t):

$$d/dt \int \varphi(x)\, p(x,t)\, dx = \int [\, f_i\, \partial\varphi/\partial x_i + (1/2)\, D_{i\square}\, \partial^2\varphi/\partial x_i \partial x \,]\, p(x,t)\, dx$$

Si integra per parti il lato destro. Per il primo termine:

$$\int f_i\, (\partial\varphi/\partial x_i)\, p\, dx = -\int \varphi\, \partial(f_i\, p)/\partial x_i\, dx + [\text{boundary terms}]$$

assumendo che p e φ decadano sufficientemente velocemente all'infinito, i termini di bordo si annullano. Per il secondo termine, integrando per parti due volte:

$$\int D_{i\square}\, (\partial^2\varphi/\partial x_i \partial x)\, p\, dx = \int \varphi\, \partial^2(D_{i\square}\, p)/\partial x_i \partial x\, dx$$

Il lato sinistro può essere scritto come:

$$d/dt \int \varphi\, p\, dx = \int \varphi\, (\partial p/\partial t)\, dx$$

Poiché l'equazione deve valere per ogni funzione test φ, si ottiene, per uguaglianza puntuale degli integrandi:

$$\partial p/\partial t = -\partial(f_i\, p)/\partial x_i + (1/2)\, \partial^2(D_{i\square}\, p)/\partial x_i \partial x$$

In notazione vettoriale compatta, questa è l'**equazione di Fokker-Planck** (o equazione di Kolmogorov in avanti):

$$\partial p/\partial t = -\nabla \cdot (f\, p) + (1/2)\, \nabla \cdot \nabla \cdot (D\, p) \quad (A.6)$$

ovvero, usando l'operatore di Laplace-Beltrami generalizzato:

$$\partial p/\partial t = -\nabla \cdot (f\, p) + (1/2)\, \Sigma_{i\square}\, \partial^2(D_{i\square}\, p)/\partial x_i \partial x$$

A.2.3 — Forma del flusso di probabilità

L'equazione (A.6) ha la struttura di un'equazione di continuità:

$$\partial p/\partial t + \nabla \cdot J = 0 \quad (A.7)$$

dove il *flusso di probabilità* J è definito da:

$$J_i = f_i\, p - (1/2)\, \Sigma\ \partial(D_i\square\ p)/\partial x \qquad \text{(A.8)}$$

Il primo termine, f p, è il flusso avvettivo dovuto al drift deterministico; il secondo termine, $-(1/2)\ \partial(D_i\square\ p)/\partial x$, è il flusso diffusivo che tende a uniformare la densità di probabilità. L'equazione di Fokker-Planck afferma che la variazione temporale della densità di probabilità in ogni punto è uguale alla divergenza negativa del flusso — il che è la condizione di conservazione della probabilità totale $\int p(x,t)\, dx = 1$ per ogni t.

A.2.4 — Il caso isotropo e il drift come gradiente

Nel caso in cui la matrice di diffusione sia isotropa e omogenea, $D(x,t) = 2D \cdot I$, l'equazione di Fokker-Planck si semplifica a:

$$\partial p/\partial t = -\nabla\cdot(f\, p) + D\ \nabla^2 p \qquad \text{(A.9)}$$

dove $\nabla^2 = \Sigma_i\ \partial^2/\partial x_i^2$ è il laplaciano scalare. Questo è il caso più direttamente rilevante per la connessione con la dinamica di Langevin nella minimizzazione dell'energia libera. Si assuma che il drift abbia la forma gradiente:

$$f(x) = -\nabla U(x) \qquad \text{(A.10)}$$

dove $U: \mathbb{R}^n \to \mathbb{R}$ è una funzione di energia potenziale (o, nel linguaggio variazionale, un'energia libera). Sostituendo in (A.9):

$$\begin{aligned}\partial p/\partial t &= \nabla\cdot(p\ \nabla U) + D\ \nabla^2 p \\ &= \nabla\cdot[\, p\ \nabla U + D\ \nabla p\,] \\ &= \nabla\cdot[\, p\ \nabla (U + D \ln p)\,] \\ &= D\ \nabla\cdot[\, p\ \nabla\ \ln(p\ e\text{^}\{U/D\})\,] \qquad \text{(A.11)}\end{aligned}$$

Questa formulazione mostra che la dinamica di Fokker-Planck con drift gradiente spinge la densità p verso la distribuzione di Gibbs-Boltzmann:

$$p^*(x) \propto \exp(-U(x)/D)$$

che è l'unico punto fisso stazionario dell'equazione (flusso $J = 0$ ovunque, condizione di bilancio dettagliato).

A.3 — La distribuzione stazionaria e la funzione di Lyapunov

A.3.1 — Caratterizzazione della distribuzione stazionaria

La distribuzione stazionaria $p^*(x)$ soddisfa $\partial p^*/\partial t = 0$, ovvero $\nabla \cdot J^* = 0$ con $J^* = f\,p^* - D\,\nabla p^*$. La condizione più forte — il *bilancio dettagliato* — richiede $J^* = 0$ identicamente, il che implica:

$$f\,p^* = D\,\nabla p^*$$
$$f = D\,\nabla \ln p^* \qquad \text{(A.12)}$$

Invertendo, se il drift ha forma gradiente $f = -\nabla U$, la condizione (A.12) è soddisfatta con:

$$\ln p^*(x) = -U(x)/D + \text{const}$$
$$p^*(x) = Z^{-1} \exp(-U(x)/D) \qquad \text{(A.13)}$$

dove $Z = \int \exp(-U(x)/D)\,dx$ è la costante di normalizzazione (funzione di partizione). Per $D = 1$ (temperature unitaria), la distribuzione stazionaria è semplicemente:

$$p^*(x) = Z^{-1} \exp(-U(x)) = \exp(-(U(x) - \ln Z))$$

Il potenziale di Gibbs $U(x)$ coincide con l'energia libera per particella del sistema fisico corrispondente.

A.3.2 — L'energia libera come funzione di Lyapunov

Si definisce il *funzionale di energia libera* relativo alla distribuzione stazionaria p^*:

$$F[p] = \int p(x)\,[\,U(x) + \ln p(x)\,]\,dx$$
$$= E_p[U] - H[p] \qquad \text{(A.14)}$$

dove $H[p] = -\int p \ln p\,dx$ è l'entropia di Shannon della distribuzione p. Questo funzionale può essere riscritto usando la distribuzione stazionaria $p^* = Z^{-1} e^{\{-U\}}$:

$$F[p] = \int p\,[\,-\ln p^* - \ln Z + \ln p\,]\,dx$$
$$= D_KL(p \,\|\, p^*) - \ln Z \qquad \text{(A.15)}$$

Poiché $D_KL(p \,\|\, p^*) \geq 0$ con uguaglianza sse $p = p^*$, si ha $F[p] \geq -\ln Z$, con minimo raggiunto in $p = p^*$.

Si calcola ora la variazione temporale di $F[p]$ lungo le traiettorie della Fokker-Planck:

$$dF/dt = \int (\partial p/\partial t)\,[\,U + \ln p + 1\,]\,dx$$
$$= \int (\partial p/\partial t)\,[\,U + \ln p\,]\,dx \qquad (\text{poiché } \int \partial p/\partial t\,dx = 0)$$

Sostituendo l'equazione di Fokker-Planck (A.9) con D=1 e $f = -\nabla U$:

$$\partial p/\partial t = \nabla \cdot (p\, \nabla U) + \nabla^2 p = \nabla \cdot [\, p(\nabla U + \nabla \ln p)\,] = \nabla \cdot [\, p\, \nabla (U + \ln p)\,]$$

Si ottiene, integrando per parti e annullando i termini di bordo:

$$dF/dt = \int \nabla \cdot [\, p\, \nabla (U + \ln p)\,] \cdot (U + \ln p)\, dx$$
$$= -\int p \, \| \nabla (U + \ln p) \|^2 \, dx \qquad \text{(A.16)}$$

Poiché l'integrando è non-negativo, si ha $dF/dt \leq 0$, con uguaglianza sse $\nabla (U + \ln p) = 0$ quasi ovunque, ossia sse $p = p^*$. L'energia libera variazionale F[p] è quindi una **funzione di Lyapunov** per la dinamica di Fokker-Planck: essa decresce monotonicamente lungo ogni traiettoria e raggiunge il suo minimo globale unicamente nella distribuzione stazionaria. Questo è il contenuto preciso dell'affermazione che i sistemi biologici "minimizzano l'energia libera": la dinamica stocastica che li descrive ha questa proprietà per costruzione, nella misura in cui la distribuzione di probabilità sui loro stati converge verso la distribuzione stazionaria p^*.

La velocità di decrescita di F, equazione (A.16), è l'*informazione di Fisher* della distribuzione p rispetto al gradiente del potenziale effettivo $U + \ln p$ — una quantità che misura la "forza" con cui la distribuzione si discosta dall'equilibrio stazionario.

A.4 — Dualità con l'equazione di Schrödinger

A.4.1 — La trasformazione di Doob

La connessione tra l'equazione di Fokker-Planck e quella di Schrödinger si stabilisce mediante la *trasformazione di Doob* — una riscrittura della densità di probabilità in termini di un'ampiezza reale positiva. Si definisce:

$$p(x, t) = \psi(x, t)^2 \qquad \text{(A.17)}$$

ovvero $\psi(x,t) = \sqrt{p(x,t)}$ — la radice quadrata della densità di probabilità (un'ampiezza reale e positiva, a differenza dell'ampiezza complessa della meccanica quantistica). Sostituendo nell'equazione di Fokker-Planck (A.9) con D=1 e $f = -\nabla U$:

$$\partial(\psi^2)/\partial t = \nabla \cdot (\psi^2\, \nabla U) + \nabla^2(\psi^2)$$

Usando $\partial(\psi^2)/\partial t = 2\psi\, \partial\psi/\partial t$ e $\nabla^2(\psi^2) = 2\psi\, \nabla^2\psi + 2\, \|\nabla\psi\|^2$:

$$2\psi\, \partial\psi/\partial t = \psi^2\, \Delta U + 2\psi\, \nabla\psi \cdot \nabla U + 2\psi\, \nabla^2\psi + 2\|\nabla\psi\|^2$$

Dividendo per 2ψ (dove ψ > 0):

$$\partial\psi/\partial t = (\psi/2)\,\Delta U + \nabla\psi \cdot \nabla U + \nabla^2\psi + \|\nabla\psi\|^2/\psi \qquad \text{(A.18)}$$

Questo può essere riscritto in forma più compatta notando che:

$$\nabla^2\psi + \nabla\psi \cdot \nabla U + (\psi/2)\,\Delta U = \nabla^2\psi + \nabla\cdot(\psi\,\nabla U) - \psi\,\|\nabla U\|^2/2 + (\psi/2)\,\Delta U$$

In realtà la forma più illuminante si ottiene passando direttamente alla sostituzione ψ = e^{-φ/2} per un opportuno potenziale φ. Si pone:

$$p(x) = Z^{-1}\, e^{\{-U(x)\}} \;\equiv\; e^{\{-\varphi(x)\}}$$

con φ(x) = U(x) + ln Z. Allora $\psi_0(x) = e^{\{-\varphi(x)/2\}} = (Z)^{\{-1/2\}}\, e^{\{-U(x)/2\}}$ è la radice quadrata della distribuzione stazionaria. Si decompone la soluzione temporale come:

$$\psi(x, t) = \psi_0(x) \cdot \chi(x, t) \qquad \text{(A.19)}$$

Sostituendo nell'equazione di Fokker-Planck riscritta per ψ e dopo calcoli algebrici — che si riportano di seguito — si ottiene per χ un'equazione di diffusione pura.

A.4.2 — L'operatore di Fokker-Planck e il suo aggiunto

L'equazione di Fokker-Planck per il caso in esame è:

$$\partial p/\partial t = \mathcal{L}_{FP}\, p \qquad \text{(A.20)}$$

dove l'operatore di Fokker-Planck è:

$$\mathcal{L}_{FP} = -\nabla\cdot(f\,\cdot) + \nabla^2\cdot \quad \text{con} \quad f = -\nabla U$$

L'*operatore aggiunto* rispetto al prodotto scalare $L^2(\mathbb{R}^n)$ è:

$$\mathcal{L}_{FP}^{\dagger} = f\cdot\nabla + \nabla^2\cdot = -\nabla U\cdot\nabla + \nabla^2$$

che è l'operatore di generatore del processo stocastico, o *operatore di Kolmogorov all'indietro*. L'operatore aggiunto simmetrizzato (hermitiano) si ottiene mediante la trasformazione di similitudine mediata da ψ_0:

$$H_{sym} = \psi_0^{-1}\,\mathcal{L}_{FP}\,\psi_0\,\cdot = \nabla^2 - V(x) \qquad \text{(A.21)}$$

dove il potenziale effettivo V(x) è:

$$V(x) = (1/4) \| \nabla U \|^2 - (1/2) \nabla^2 U \qquad \text{(A.22)}$$

Questo si verifica esplicitamente. Si applica $\psi_0^{-1}\, \mathcal{L}_FP\, (\psi_0\, \chi)$ per una funzione arbitraria χ:

$$\begin{aligned} \mathcal{L}_FP(\psi_0\, \chi) &= -\nabla \cdot (-\nabla U \cdot \psi_0\, \chi) + \nabla^2 (\psi_0\, \chi) \\ &= \nabla U \cdot \nabla (\psi_0\, \chi) + (\Delta U)\, \psi_0\, \chi + \nabla^2 (\psi_0\, \chi) \end{aligned}$$

Si calcolano i termini usando $\nabla \psi_0 = -(1/2)(\nabla U)\, \psi_0$:

$$\nabla (\psi_0\, \chi) = (\nabla \psi_0)\, \chi + \psi_0\, \nabla \chi = \psi_0 (-\nabla U/2 \cdot \chi + \nabla \chi)$$

$$\begin{aligned} \nabla^2 (\psi_0\, \chi) &= \nabla \cdot [\, \psi_0 (-\nabla U/2 \cdot \chi + \nabla \chi)\,] \\ &= \psi_0 [(1/4) \| \nabla U \|^2 \chi - (1/2)(\Delta U)\chi - (\nabla U/2) \cdot \nabla \chi - (\nabla U/2) \cdot \nabla \chi + \nabla^2 \chi] \\ &= \psi_0 [(1/4) \| \nabla U \|^2 - (1/2)\Delta U]\, \chi - \psi_0 (\nabla U) \cdot \nabla \chi + \psi_0\, \nabla^2 \chi \end{aligned}$$

Mettendo insieme i contributi e dividendo per ψ_0:

$$\psi_0^{-1}\, \mathcal{L}_FP(\psi_0\, \chi) = \nabla^2 \chi - V(x)\, \chi \qquad \text{(A.23)}$$

con V(x) definito come in (A.22). L'equazione per $\chi(x,t) = \psi(x,t)/\psi_0(x)$ è dunque:

$$\partial \chi / \partial t = \nabla^2 \chi - V(x)\, \chi \qquad \text{(A.24)}$$

che è un'**equazione di diffusione con potenziale** — formalmente identica a un'equazione di Schrödinger in tempo immaginario $\tau = it$, una volta identificata $\hbar = 1$ e $m = 1/2$:

$$-\partial \chi / \partial \tau = -\nabla^2 \chi + V(x)\, \chi \quad \text{ovvero} \quad \partial \chi / \partial \tau = [\, \nabla^2 - V(x)\,]\, \chi$$

A.4.3 — La rotazione di Wick e la dualità con l'equazione di Schrödinger

L'equazione di Schrödinger per una particella di massa m in un potenziale V(x) è:

$$i\hbar\, \partial \Psi / \partial t = [\, -\hbar^2/(2m)\, \nabla^2 + V(x)\,]\, \Psi \qquad \text{(A.25)}$$

La *rotazione di Wick* consiste nella sostituzione del tempo reale t con il tempo immaginario $\tau = -it$ (o equivalentemente $t = i\tau$). Sotto questa sostituzione, i $\partial/\partial t$ diviene $-\partial/\partial \tau$, e l'equazione di Schrödinger diviene:

$$\hbar\, \partial\Psi/\partial\tau = [\, -\hbar^2/(2m)\ \nabla^2 + V(x)\,]\ \Psi \qquad \text{(A.26)}$$

Per $\hbar = 1$, $m = 1/2$ (che fissa l'unità di massa) e moltiplicando per -1:

$$-\partial\Psi/\partial\tau = [\, \nabla^2 - V(x)\,]\ \Psi \quad \text{ovvero} \quad \partial\Psi/\partial\tau = -[\, \nabla^2 - V(x)\,]\ \Psi$$

Confrontando con l'equazione (A.24) per χ, si vede che:

equazione per χ: $\partial\chi/\partial t = [\nabla^2 - V(x)]\ \chi$
Schrödinger in τ: $\partial\Psi/\partial\tau = -[\nabla^2 - V(x)]\ \Psi$

Le due equazioni sono legate da $\Psi = \chi$ con $\tau = -t$ (o equivalentemente $t = -\tau$). Questa è la *dualità Fokker-Planck/Schrödinger* in tempo immaginario: il propagatore della Fokker-Planck nel tempo reale corrisponde al propagatore di Schrödinger nel tempo immaginario con il segno opposto del potenziale.

La conseguenza più rilevante di questa dualità riguarda lo spettro dell'operatore. L'operatore hermitiano $H_sym = -\nabla^2 + V(x)$ ha uno spettro di autovalori reali $\{\lambda\}$ con autofunzioni ortonormali $\{\varphi(x)\}$ che soddisfano:

$$H_sym\ \varphi = \lambda\ \varphi$$

L'autovalore minimo $\lambda_0 = 0$ corrisponde all'autofunzione fondamentale $\varphi_0(x) = \psi_0(x)$ — la radice quadrata della distribuzione stazionaria. Tutti gli altri autovalori sono positivi: $\lambda > 0$ per $n \geq 1$. La soluzione generale dell'equazione di Fokker-Planck è:

$$p(x, t) = \psi_0(x)\ [\, c_0 + \Sigma_\{n\geq 1\}\ c\ \varphi(x)/\psi_0(x) \cdot e\hat{}\{-\lambda\ t\}\,]$$

dove i coefficienti c sono determinati dalla condizione iniziale. Per $t \to \infty$, tutti i termini con $n \geq 1$ decadono esponenzialmente, e la soluzione converge verso la distribuzione stazionaria $p^*(x) \propto \psi_0(x)^2$. Il tasso di convergenza è determinato dal primo autovalore non-nullo λ_1 (il *gap spettrale*), che gioca in questo contesto il ruolo della costante di tempo di rilassamento verso l'equilibrio.

A.4.4 — Significato per il framework variazionale

La dualità appena derivata ha tre implicazioni per il corpo principale del saggio. La prima è di carattere algoritmico: gli algoritmi di campionamento stocastico come la dinamica di Langevin e la Monte Carlo hamiltoniana, usati per campionare da distribuzioni posterior complesse, sono esattamente l'equazione (A.1) con il potenziale $U = -\ln P(\psi|s)$. La loro convergenza verso la distribuzione target è garantita dalla teoria di Fokker-Planck, con tasso determinato dal gap spettrale λ_1 dell'operatore hermitiano associato. Questo fornisce criteri quantitativi per confrontare la velocità di convergenza di diversi schemi di campionamento.

La seconda implicazione riguarda il legame con la meccanica quantistica del campo. La propagazione del calore — equazione di Fokker-Planck omogenea, $V = 0$ — è duale alla propagazione quantistica libera. L'inclusione del potenziale $V(x) = (1/4)\|\nabla U\|^2 - (1/2)\Delta U$ introduce interazioni che, nel linguaggio della teoria quantistica dei campi, corrispondono a termini di massa e autointerazione. Questa connessione non è puramente formale: il formalismo del path integral, sviluppato da Feynman per la meccanica quantistica, si applica direttamente al calcolo dei propagatori di Fokker-Planck mediante la sostituzione $t \to -i\tau$, fornendo rappresentazioni integrali esatte degli stessi propagatori che nella pratica si approssimano con la variational inference.

La terza implicazione è concettuale e riguarda il punto sollevato in IV.6 sui possibili fondamenti quantistici dell'inferenza neurale. La dualità Fokker-Planck/Schrödinger non implica che il cervello "computazioni quantistiche" nel senso di Penrose-Hameroff. Implica qualcosa di più sottile: che la struttura matematica dell'inferenza ottimale in presenza di rumore — la minimizzazione dell'energia libera nella dinamica stocastica — è isomorfa alla struttura della meccanica quantistica in tempo immaginario. Questo isomorfismo suggerisce che certi algoritmi di inferenza variazionale potrebbero beneficiare di implementazioni in sistemi fisici che sfruttano naturalmente la struttura del propagatore quantistico — hardware quantistico per inferenza bayesiana — non perché il cervello sia quantistico, ma perché la struttura matematica del problema lo è.

Note Bibliografiche — Fonti Primarie, Letteratura Tecnica e Orientamenti Critici

---

Le note che seguono non costituiscono un elenco esaustivo della letteratura sul tema, né si limitano a segnalare le fonti da cui questo saggio ha attinto direttamente. Il loro scopo è orientare il lettore entro una letteratura straordinariamente dispersa — che attraversa neuroscienze computazionali, teoria dell'informazione, geometria differenziale e filosofia della mente — indicando le opere fondative, le controversie aperte e i percorsi di approfondimento più produttivi per ciascuno dei temi trattati.

Radici storiche ed epistemologiche

L'opera di Hermann von Helmholtz che fonda l'intera tradizione è il *Handbuch der physiologischen Optik* (tre volumi, 1856–1867), di cui esiste una traduzione inglese curata da James Southall per la Optical Society of America (1924, ristampata da Dover nel 1962). Il concetto di *unbewusster Schluss* è sviluppato soprattutto nel terzo volume, nelle sezioni dedicate alla percezione dello spazio. Per una lettura filosoficamente attenta di Helmholtz in rapporto a Kant, il testo di riferimento è Gary Hatfield, *The Natural and the Normative: Theories of Spatial Perception from Kant to Helmholtz* (MIT Press, 1990), che

argomenta come la "naturalizzazione" helmholtziana del kantismo sia già una forma di epistemologia naturalizzata nel senso quiniano.

La proposta di Richard Gregory sulla percezione come formulazione di ipotesi è esposta in modo accessibile in *Eye and Brain: The Psychology of Seeing* (McGraw-Hill, 1966; quinta edizione Weidenfeld and Nicolson, 1998). Per l'elaborazione più filosoficamente rigorosa: Gregory, "Perceptions as hypotheses", *Philosophical Transactions of the Royal Society B*, 290 (1980), 181–197. Gregory rimane un interlocutore trascurato nella letteratura tecnica contemporanea, e la sua riscoperta nel contesto della codifica predittiva è uno dei contributi intellettualmente più onesti di Jakob Hohwy, *The Predictive Mind* (Oxford University Press, 2013) — il testo filosoficamente più rigoroso sull'argomento disponibile in lingua inglese.

Inferenza bayesiana e cervello probabilistico

Il programma del "cervello bayesiano" nelle sue formulazioni quantitative fondamentali è documentato da tre lavori di riferimento. Weiss, Simoncelli e Adelson, "Motion illusions as optimal percepts", *Nature Neuroscience* 5 (2002), 598–604, è il testo paradigmatico che mostra come le illusioni visive di movimento siano predizioni bayesiane ottimali sotto priori sulla lentezza dei movimenti naturali. Kersten, Mamassian e Yuille, "Object perception as Bayesian inference", *Annual Review of Psychology* 55 (2004), 271–304, è la revisione sistematica più citata del programma. Knill e Pouget, "The Bayesian brain: the role of uncertainty in neural coding and computation", *Trends in Neurosciences* 27 (2004), 712–719, offre la prospettiva teorica più esplicita sulla codifica della distribuzione a posteriori nei pattern di attività neurale.

Il problema dell'implementazione neurale dell'inferenza bayesiana è affrontato da Pouget, Beck, Ma e Latham, "Probabilistic brains: knowns and unknowns", *Nature Neuroscience* 16 (2013), 1170–1178, che distingue le rappresentazioni implicite mediante campionamento da quelle esplicite mediante distributional coding e discute l'evidenza empirica a favore di ciascuna.

Inferenza variazionale: fondamenti matematici

La derivazione dell'ELBO e le proprietà fondamentali dell'inferenza variazionale sono trattate con completezza matematica in David Blei, Alp Kucukelbir e Jon McAuliffe, "Variational Inference: A Review for Statisticians", *Journal of the American Statistical Association* 112 (2017), 859–877. Questo articolo è la referenza tecnica più accessibile e completa per chi affronta il soggetto dal lato statistico. Per la prospettiva della probabilità bayesiana pura, il testo classico è Christopher Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006), capitoli 9–10, che tratta l'EM e l'inferenza variazionale per i modelli grafici.

La teoria della famiglia esponenziale e la relazione tra inferenza variazionale e massima entropia è sviluppata rigorosamente in Wainwright e Jordan, *Graphical Models, Exponential Families, and Variational Inference* (Foundations and Trends in Machine Learning, 2008) — il testo tecnico di riferimento per i metodi variazionali moderni, inclusi quelli su cui si basa il machine learning contemporaneo.

Geometria dell'informazione

L'opera fondativa è Shun-ichi Amari, *Information Geometry and Its Applications* (Springer, 2016), che raccoglie e sistematizza quattro decenni di ricerca. Per un'introduzione più accessibile che mantenga il rigore: Amari e Nagaoka, *Methods of Information Geometry* (American Mathematical Society / Oxford University Press, 2000), traduzione dal giapponese. Il teorema di Chentsov sull'unicità della metrica di Fisher è dimostrato in Chentsov, *Statistical Decision Rules and Optimal Inference* (American Mathematical Society, 1982, traduzione dal russo del 1972).

La connessione tra gradiente naturale e apprendimento è stata stabilita da Amari, "Natural Gradient Works Efficiently in Learning", *Neural Computation* 10 (1998), 251–276 — un articolo che ha avuto enorme influenza nell'ottimizzazione dei modelli di apprendimento automatico e che anticipa di quasi vent'anni molte idee dei metodi di ottimizzazione per reti neurali profonde.

La geometria del trasporto ottimale e la metrica di Wasserstein sono trattate in modo accessibile e completo in Cédric Villani, *Optimal Transport: Old and New* (Springer, 2009). Il risultato di Jordan-Kinderlehrer-Otto sulla discesa di gradiente di Wasserstein è in Jordan, Kinderlehrer e Otto, "The variational formulation of the Fokker-Planck equation", *SIAM Journal of Mathematical Analysis* 29 (1998), 1–17.

Principio di energia libera e inferenza attiva

La serie di articoli fondativi di Karl Friston sul principio di energia libera comprende: Friston, "A free energy principle for a particular physics", *arXiv:1906.10184* (2019), che è il testo più completo e matematicamente esplicito; Friston, "The free-energy principle: a unified brain theory?", *Nature Reviews Neuroscience* 11 (2010), 127–138, che è la formulazione più citata e accessibile; Friston, Kilner e Harrison, "A free energy principle for the brain", *Journal of Physiology Paris* 100 (2006), 70–87, che è il testo di fondazione originale.

Per l'inferenza attiva come estensione al comportamento: Friston et al., "Active inference and epistemic value", *Cognitive Neuroscience* 6 (2015), 187–214. La distinzione tra valore pragmatico e valore epistemico nell'energia libera attesa è qui elaborata con dettaglio formale. Un libro recente che raccoglie l'intera teoria in forma pedagogica è Parr, Pezzulo e Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and*

Behavior (MIT Press, 2022), il testo di riferimento per chi voglia padroneggiare il framework senza passare attraverso la dispersa letteratura originale.

La struttura del Markov blanket nel contesto della teoria dei sistemi complessi è discussa in Friston, "Life as we know it", *Journal of the Royal Society Interface* 10 (2013), 20130475. Per una critica epistemologica della portata del principio: Colombo e Series, "Bayes in the Brain — On Bayesian Modelling in Neuroscience", *The British Journal for the Philosophy of Science* 63 (2012), 697–723.

Codifica predittiva gerarchica

Il modello originale di Rao e Ballard è in Rao e Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects", *Nature Neuroscience* 2 (1999), 79–87. Questo articolo è breve ma denso, e la derivazione matematica presentata nel corpo principale del saggio è più esplicita di quella nell'originale. Il lavoro di Mumford, "On the computational architecture of the neocortex. II. The role of corticocortical loops", *Biological Cybernetics* 66 (1992), 241–251, anticipa l'idea senza il formalismo.

La connessione tra oscillazioni corticali e segnali di codifica predittiva è sviluppata in Bastos et al., "Canonical microcircuits for predictive coding", *Neuron* 76 (2012), 695–711 — l'articolo che ha sistematizzato la proposta anatomo-funzionale e ha stimolato una generazione di studi di connettività spettrale. Per le prove elettrofisiologiche nelle bande gamma e alpha-beta: Michalareas et al., "Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas", *Neuron* 89 (2016), 384–397.

L'attenzione come inferenza sulla precisione è trattata da Feldman e Friston, "Attention, uncertainty, and free-energy", *Frontiers in Human Neuroscience* 4 (2010), 215. L'architettura PredNet ispirata alla codifica predittiva per l'apprendimento video è in Lotter, Kreiman e Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning", *ICLR* 2017.

Psicopatologia e neuroscienze computazionali

Il framework dell'aberrant salience per la schizofrenia in termini bayesiani è proposto da Fletcher e Frith, "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia", *Nature Reviews Neuroscience* 10 (2009), 48–58. L'evidenza sperimentale basata sulla mismatch negativity (MMN) è raccolta in Näätänen et al., "Mismatch negativity (MMN) as an index of cognitive dysfunction", *Brain Topography* 27 (2014), 451–466.

Per l'autismo come condizione di priori iperpredicibili: Pellicano e Burr, "When the world becomes 'too real': a Bayesian explanation of autistic perception", *Trends in*

Cognitive Sciences 16 (2012), 504–510. La formalizzazione in termini di inferenza variazionale con alterata precisione è in Lawson, Rees e Friston, "An aberrant precision account of autism", *Frontiers in Human Neuroscience* 8 (2014), 302.

Filosofia della mente: realismo, enattivismo, coscienza

Per il realismo strutturale: John Worrall, "Structural Realism: The Best of Both Worlds?", *Dialectica* 43 (1989), 99–124 è il testo fondante; James Ladyman e Don Ross, *Every Thing Must Go: Metaphysics Naturalized* (Oxford University Press, 2007) è la sua elaborazione più radicale e connessa alle scienze speciali. Il legame con le neuroscienze computazionali è esplorato in Michael Rescorla, "The Representational Foundations of Computation", *Philosophia Mathematica* 23 (2015), 338–366.

Per l'enattivismo: Maturana e Varela, *The Tree of Knowledge* (Shambhala, 1987) e Alva Noë, *Action in Perception* (MIT Press, 2004) sono le fonti primarie. Il tentativo di riconciliazione con la codifica predittiva è in Hohwy, "The self-evidencing brain", *Noûs* 50 (2016), 259–285, e da una prospettiva più critica verso Hohwy, in Evan Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Harvard University Press, 2007).

Il problema difficile nella sua formulazione canonica è in David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, 1996). Il panpsichismo strutturale è difeso in Phillip Goff, *Consciousness and Fundamental Reality* (Oxford University Press, 2017). La teoria dell'informazione integrata è formulata in Tononi, "Consciousness as Integrated Information: a Provisional Manifesto", *Biological Bulletin* 215 (2008), 216–242. Una critica rigorosa e misurata dell'IIT è in Scott Aaronson, "Why I Am Not An Integrated Information Theorist", blog post (2014), accessibile all'indirizzo scottaaronson.blog — testo informale ma di rara lucidità tecnica sul problema della falsificabilità dell'IIT.

Equazione di Fokker-Planck e meccanica stocastica

La derivazione standard dell'equazione di Fokker-Planck per processi di Itô è in Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer, quarta edizione 2009), capitoli 3–4 — il testo tecnico di riferimento assoluto per il tema. La connessione con la meccanica quantistica in tempo immaginario mediante rotazione di Wick è discussa in Parisi e Sourlas, "Random Magnetic Fields, Supersymmetry, and Negative Dimensions", *Physical Review Letters* 43 (1979), 744–745 e, in contesto più accessibile, in Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Oxford University Press, quarta edizione 2002), capitolo 4. Il risultato di Jordan-Kinderlehrer-Otto citato nel testo è in Jordan, Kinderlehrer e Otto, "The variational formulation of the Fokker-Planck equation", *SIAM Journal of Mathematical Analysis* 29 (1998), 1–17.